Introduction to Natural Language Processing
Assignment 6. Get started on question answering. Due Next Wed at midnight
The next three homework assignments will be to design and build a question answering system. We will use stories from Aesop’s Fables and the Blogs Corpus. For each homework you will receive new datasets with more difficult stories and questions.

We created questions and an answer key for each story. Each question has a unique \textit{QuestionID} which is the \textit{StoryID} followed by a dash and question number. For example, “blogs-01-1” means that this is question #1 pertaining to story “blogs-01”. In some cases, the answer key allows for several acceptable answers (e.g., “Toronto, Ontario” vs. “Toronto”), paraphrases (e.g., “Human Immunodeficiency Virus” vs. “HIV”), varying amounts of information (e.g., “he died” vs. “he died in his sleep of natural causes”), or occasionally different interpretations of the question (e.g., “Where did the boys learn how to survive a storm?” “camping tips from a friend” vs. “their backyard”). When more than one answer is acceptable, the acceptable answers are separated by a vertical bar (|). Below is a sample answer key:

<table>
<thead>
<tr>
<th>QuestionID: blogs-01-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question: What is the summit meeting named?</td>
</tr>
<tr>
<td>Answer: G20 summit</td>
</tr>
<tr>
<td>Difficulty: Easy</td>
</tr>
<tr>
<td>Type: Sch</td>
</tr>
</tbody>
</table>
How is it going to be set up?

blogs-01.answers  Shared

blogs-01.questions  Shared

blogs-01.questions.dep  Shared

blogs-01.questions.par  Shared

blogs-01.sch  Shared

blogs-01.sch.dep  Shared

blogs-01.sch.par  Shared
When did the G20 summit start? (sch)
What is the summit meeting named? (sch) pos np
Where did the protest happen? (sch | story) loc NE regex
on a street | along the street where I work | right in front of my store
What did the people burn? (sch)
Who rebelled? (sch)
Who created a riot? (sch)
What happened to a police car? (sch)
Who burned a police car? (sch)
What was burned? (sch | story)
What happened to the police cars? (story)
What was fired? (story)
What was fired at the rioters? (story)
What was smashed? (story)
QuestionID: blogs-01-1
Question: When did the G20 summit start?
Answer: on eventful today | eventful today
Difficulty: Easy
Type: Sch

QuestionID: blogs-01-2
Question: What is the summit meeting named?
Answer: G20 summit
Difficulty: Easy
Type: Sch

QuestionID: blogs-01-3
Question: Where did the protest happen?
Answer: on a street | along the street where I work | right in front of my store
Difficulty: Easy
Type: Story | Sch

QuestionID: blogs-01-4
Question: What did the people burn?
Answer: a police car | police car
Difficulty: Easy
Type: Sch

QuestionID: blogs-01-5
Question: Who rebelled?
Difficulty: Easy
Type: Sch

QuestionID: blogs-01-6
Question: Who created a riot?
Difficulty: Easy
Today was a very eventful work day. Today was the start of the G20 summit. It happens every year and it is where 20 of the leaders of the world come together to talk about how to run their governments effectively and what not. Since there are so many leaders coming together their are going to be a lot of people who have different views on how to run the government they follow so they protest. There was a protest that happened along the street where I work and at first it looked peaceful until a bunch of people started rebelling and creating a riot. Police cars were burned and things were thrown at cops. Police were in full riot gear to alleviate the violence. As things got worse tear gas and bean bag bullets were fired at the rioters while they smash windows of stores. And this all happened right in front of my store which was kind of scary but it was kind of interesting since I've never seen a riot before.
A summit meeting named G20 summit started on eventful today.

G20 summit happened annually.

A world and many leader came and talked about it running a government.

A people protested because it disagreed about a view.

The people protested peacefully on a street.

The people rebelled and created riot.

The people burned a police car and threw a thing at a police.

The police alleviated the people of riot.

The police fired a tear gas at the people and fired a bullet at the people, and the people smashed a window.
Sch realization resolves pronouns, but may change words
Okay what are all these other files?
Dep Pareses for the questions and stories

- USE the TREE READER to convert
Constituent Parses .par files

- USE the TREE READER to convert

QuestionId: blogs-01-1
(ROOT (SBARQ (WHADVP (WRB When))) (SQ (VBD did) (NP (DT the) (NN G20) (NN summit))) (VP (VB start))) (. ?)))

QuestionId: blogs-01-10
(ROOT (SBARQ (WHNP (WP What))) (SQ (VP (VBD happened) (PP (TO to) (NP (DT the) (NN police) (NNS cars)))))) (. ?)))

QuestionId: blogs-01-11
(ROOT (SBARQ (WHNP (WP What))) (SQ (VBD was) (VP (VBN fired)))) (. ?)))
HW 6

- TWO Fables for training with questions and answers
- ONE BLOG for training with questions and answers
- In one week, next Wed, we will test your program on a DEVELOPMENT SET, expanded set of questions of new type
- HW 7: GIVE you the DEVELOPMENT set we used to test
  - GIVE you a NEW training set with different additional kinds of questions
  - TEST AGAIN a week later (May 28th due date midnight)
- HW8: Final TEST run on June 4th (turn in June 3rd)
- Class presentations June 5th.
Performance Evaluation

2 main requirements:

- **Evaluation metric**: mathematically defines how to measure the system’s performance against human-annotated gold standard

- **Scoring program**: implements the metric and provides performance measures
  - For each question and over all questions
  - Can be easily run at any time to continually test your system performance and see what any changes have improved or worsened
  - Regression testing: Sets of questions whose performance should be maintained
Evaluation Metrics

- **Precision** = correct answers/answers produced (what proportion of the answers produced are accurate?)

- **Recall** = correct answers/total possible correct answers (what proportion of all the correct answers did the system find?)

- Trade-off between Precision and Recall
- **F1** (balanced) Measure = \( \frac{2PR}{2(R + P)} \)
**Precision and Recall**

- **Precision**: fraction of answers that are correct =
  - \( P(\text{correct} | \text{in\_the\_answer}) \)

- **Recall**: fraction of answers that are returned vs. all answers in the text =
  - \( P(\text{in\_the\_answer} | \text{correct}) \)

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>Not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the answer</td>
<td>( tp )</td>
<td>( fp )</td>
</tr>
<tr>
<td>Not in the answer</td>
<td>( fn )</td>
<td>( tn )</td>
</tr>
</tbody>
</table>

Precision \( P = \frac{tp}{(tp + fp)} \)  TRUE POSITIVES % of POSITIVES

Recall \( R = \frac{tp}{(tp + fn)} \)  TRUE POSITIVES % ANSWERS
Recall vs. Precision

- How would you get perfect precision?

- How would you get perfect recall?
Recall vs. Precision

- High recall:
  - You get all the right answers, but garbage too.
  - Good when incorrect results are not problematic.
  - More common from statistical systems

- High precision:
  - When all returned answers must be correct.
  - Good when missing results are not problematic.
  - More common from hand-built systems.

- Harder to score well on both
Evaluating Answers

- There are several ways to present an answer to a user
- Short Answer: the exact answer to the question
- Answer Sentence: the sentence containing the answer
- Answer Passage: A short passage containing the answer
Evaluating Answers

Short answers are difficult to score automatically because many variations are often acceptable.

Text: The 2002 Winter Olympics will be held in beautiful Salt Lake City, Utah.

Q: Where will the 2002 Winter Olympics be held?

• A1: beautiful Salt Lake City, Utah
• A2: Salt Lake City, Utah
• A3: Salt Lake City,
• A4: Salt Lake
• A25: Utah
Our Scoring Program: partial credit

- **Precision**: fraction of answers (question answer words) that are correct = $P(\text{correct} | \text{in_the_answer})$

- **Recall**: fraction of right words that are in_the_answer vs. all words in the answer = $P(\text{in_the_answer} | \text{correct})$

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>Not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In the answer</strong></td>
<td>$tp$</td>
<td>$fp$</td>
</tr>
<tr>
<td><strong>Not in the answer</strong></td>
<td>$fn$</td>
<td>$tn$</td>
</tr>
</tbody>
</table>
DEMO SCORING Program
Recall vs. Precision

- How would you get perfect precision?

- How would you get perfect recall?
“We didn’t underperform. You overexpected.”
How to get started: Question Answering Pipeline
Types of Question Answering


- Factoid
  - Who discovered oxygen?
  - When did Hawaii become a state?
  - Where is Ayers Rock?
  - What team won the World Series in 1992?

- List
  - What countries export oil?
  - Name U.S. cities that have a “Shubert” theater.

- Definition
  - Who is Aaron Coepland?
  - What is a quasar?
<table>
<thead>
<tr>
<th>QuestionID: blogs-01-1</th>
<th>Question: When did the G20 summit start?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty: Easy</td>
<td>Answer: on eventful today</td>
</tr>
<tr>
<td>Type: Sch</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QuestionID: blogs-01-2</th>
<th>Question: What is the summit meeting named?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty: Easy</td>
<td>Answer: G20 summit</td>
</tr>
<tr>
<td>Type: Sch</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QuestionID: blogs-01-3</th>
<th>Question: Where did the protest happen?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty: Easy</td>
<td>Answer: on a street</td>
</tr>
<tr>
<td>Type: Story</td>
<td>Sch</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QuestionID: blogs-01-4</th>
<th>Question: What did the people burn?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty: Easy</td>
<td>Answer: a police car</td>
</tr>
<tr>
<td>Type: Sch</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QuestionID: blogs-01-5</th>
<th>Question: Who rebelled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty: Easy</td>
<td></td>
</tr>
<tr>
<td>Type: Sch</td>
<td></td>
</tr>
</tbody>
</table>
Generic QA Architecture

Architecture of Typical Q/A System

- Question
- Text(s)
- Question Typing
- Document/Passage Retrieval
- Named Entity Tagging
- Answer Identification

entity type(s)
relevant text(s)
tagged text
Central Idea of Factoid QA

- Determine the semantic type of the expected answer

  “Who won the Nobel Peace Prize in 1991?” is looking for a PERSON

- Retrieve documents that have keywords from the question

  Retrieve documents that have the keywords “won”, “Nobel Peace Prize”, and “1991”

- Look for named-entities of the proper type near keywords

  Look for a PERSON near the keywords “won”, “Nobel Peace Prize”, and “1991”
An Example

Who won the Nobel Peace Prize in 1991?

But many foreign investors remain sceptical, and western governments are withholding aid because of the Slorc's dismal human rights record and the continued detention of Ms Aung San Suu Kyi, the opposition leader who won the Nobel Peace Prize in 1991.

The military junta took power in 1988 as pro-democracy demonstrations were sweeping the country. It held elections in 1990, but has ignored their result. It has kept the 1991 Nobel peace prize winner, Aung San Suu Kyi - leader of the opposition party which won a landslide victory in the poll - under house arrest since July 1989.

The regime, which is also engaged in a battle with insurgents near its eastern border with Thailand, ignored a 1990 election victory by an opposition party and is detaining its leader, Ms Aung San Suu Kyi, who was awarded the 1991 Nobel Peace Prize. According to the British Red Cross, 5,000 or more refugees, mainly the elderly and women and children, are crossing into Bangladesh each day.
In OUR CASE

- We still want to determine the semantic type of the expected answer
  “Where did the protest happen?” “Where was the crow”
- Is looking for something that can specify a LOCATION
- Not just named entities. Prepositional PHRASES with LOCATIVE PREPOSITIONS (in | along| on| under | near | at | to | in front of....)

Look for sentences that include a location near the word “protest” or “crow”

“a protest that happened along the street where I work”
“A Crow was sitting on the branch of a tree.”
Generic QA Architecture

Architecture of Typical Q/A System

Question

Text(s)

Question Typing

Document/Passage Retrieval

entity type(s)

relevant text(s)

Named Entity Tagging

tagged text

Answer Identification
In OUR CASE

Don’t have to do document search. You are given the document, already processed in different forms.

• But need to retrieve passages. Zero in on likely answers.
• Rank SENTENCES within the story
• Several possible ways to rank
  • Simplest: COUNT THE OVERLAPPING WORDS of the question and the passage
  • Higher Precision:
    • Use the sentence PARSE (given in the data) and look for matching trees.
    • Extract Dependencies: look for dependency relation matches
Question TYPES

- Question word cues
  - Who → person, organization, location (e.g., city)
  - When → date
  - Where → location
  - What/Why/How → ??

- Head noun cues
  - What city, which country, what year...
  - Which astronaut, what blues band, ...

- Scalar adjective cues
  - How long, how fast, how far, how old, ...

- Answer Types can be learned or based on rules
We suggest you write a set of rules using regular expressions

- Look at the types of question you are getting and imagine the most general case of the question
- Write rules to pinpoint the type of the answer the question is asking for
- See if you can define ways to identify answers of that type

- Where is A => A is {in | under | at | along | around...}
Question Reformulation

• Question can be reformulated or expanded to improve the odds of being able to find the matching context (passage)

• Question expansion might add similar words using WordNet or similar verbs using VerbNet

• Question reformulation rules rewrite the question as a declarative statement (better match the text), e.g.
  - Where did the protest happen? => The protest happened <LOC-PHRASE>

• Sample rules from Lin 2007
  - Wh-word did A verb B => A verb+ed B
  - Where is A => A is {in | under | at | along | around...}
Picking the best answer

**Word Overlap**

A simple method for Answer Identification is to measure the amount of **Word Overlap** between the question and an answer candidate.

**Basic Word Overlap:** Each answer candidate is scored by counting how many question words are present in or near the candidate.

**Stop Words:** sometimes closed class words (often called *Stop Words* in IR) are not included in the word overlap measure.

**Roots:** sometimes stemming or morphological analysis is used to match the root forms of words (e.g., “walk” and “walked” would match).

**Weights:** some words may be weighted more heavily than others (e.g., verbs might be given more weight than nouns)
Possible to LEARN patterns from examples

Examples of Learned Patterns

**BIRTHYEAR**
*When was X born?*

<table>
<thead>
<tr>
<th>Prec</th>
<th>Pattern</th>
<th>Prec</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>(&lt;NAME&gt;) ((&lt;ANS&gt; -))</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;).</td>
</tr>
<tr>
<td>.85</td>
<td>(&lt;NAME&gt;) was born on (&lt;ANS&gt;),</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;).</td>
</tr>
<tr>
<td>.60</td>
<td>(&lt;NAME&gt;) was born in (&lt;ANS&gt;)</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.59</td>
<td>(&lt;NAME&gt;) was born (&lt;ANS&gt;)</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.53</td>
<td>(&lt;ANS&gt;) (&lt;NAME&gt;) was born</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.50</td>
<td>(-((&lt;NAME&gt;) (&lt;ANS&gt;))</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.36</td>
<td>(&lt;NAME&gt;) (((&lt;ANS&gt;) -))</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.32</td>
<td>(&lt;NAME&gt;) (((&lt;ANS&gt;)))</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.28</td>
<td>born in (&lt;ANS&gt;), (&lt;NAME&gt;)</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.20</td>
<td>of (&lt;NAME&gt;) (((&lt;ANS&gt;)))</td>
<td>.96</td>
<td>the (&lt;NAME&gt;) in (&lt;ANS&gt;)</td>
</tr>
</tbody>
</table>

**LOCATION**
*Where is X located?*

<table>
<thead>
<tr>
<th>Prec</th>
<th>Pattern</th>
<th>Prec</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;).</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;).</td>
</tr>
<tr>
<td>1.0</td>
<td>regional: (&lt;ANS&gt;): (&lt;NAME&gt;)</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;)</td>
</tr>
<tr>
<td>1.0</td>
<td>to (&lt;ANS&gt;)'s (&lt;NAME&gt;)</td>
<td>1.0</td>
<td>(&lt;ANS&gt;)'s (&lt;NAME&gt;) in</td>
</tr>
<tr>
<td>.96</td>
<td>the (&lt;NAME&gt;) in (&lt;ANS&gt;)</td>
<td>.96</td>
<td>the (&lt;NAME&gt;) in (&lt;ANS&gt;)</td>
</tr>
<tr>
<td>.92</td>
<td>from (&lt;ANS&gt;)'s (&lt;NAME&gt;)</td>
<td>.92</td>
<td>from (&lt;ANS&gt;)'s (&lt;NAME&gt;)</td>
</tr>
<tr>
<td>.92</td>
<td>near (&lt;NAME&gt;) in (&lt;ANS&gt;)</td>
<td>.92</td>
<td>near (&lt;NAME&gt;) in (&lt;ANS&gt;)</td>
</tr>
</tbody>
</table>
And there is a lot of data to learn from