Question Answering

Paola Velardi, Johan Bos
Outline

• Introduction: History of QA; Architecture of a QA system; Evaluation.

• Question Classification: NLP techniques for question analysis;

• Document Analysis: Syntactic & Semantic analysis; WordNet and other NLP resources

• Retrieving Answers: Matching; Use of Knowledge sources; Reranking; Sanity checking.
What is Question Answering?
What is QA?

Information required:

*Average number of car accidents per year in Sweden.*

Two ways of getting this information:

- Ask Google or a similar search engine (good luck!)
- Ask a QA system the question: *What’s the rate of car accidents in Sweden?*
QA vs IR

• Traditional method for information access: IR (Information Retrieval)

  – Think of IR as finding the “right book in a library”

  – Think of QA as a “librarian giving you the book and opening it on the page with the information you’re looking for”
QA vs IE

• Traditional method for information access: IE (Information Extraction)

  – Think of IE as finding answers to a pre-defined question (i.e., a template)

  – Think of QA as asking any question you like
What is Question Answering?

• Questions in natural language, not queries!

• Answers, not documents!
Why do we need QA?

• Accessing information using traditional methods such as IR and IE are limited

• Examples in the past lesson:
  – Handling negation (painkillers that do NOT cause stomach upset)
  – Connecting facts together (how many world cups have been disputed in South America?)
  – Expanding with hypernims/synonyms (tylonor is a painkiller)
  – etc
Why QA is increasingly important

• QA increasingly important because:
  – Size of available information grows
  – There is duplicate information
  – There is false information
  – People want specific information
  – More and more “computer illiterates” accessing electronically stored information
Why is QA hard? (1/3)

• **Questions** are expressed in natural language (such as English or Italian)

• Unlike formal languages, natural languages allow a great deal of flexibility

• Example:
  – What is the population of Rome?
  – How many people live in Rome?
  – What’s the size of Rome?
  – How many inhabitants does Rome have?
Why is QA hard? (2/3)

• **Answers** are expressed in natural language (such as English or Italian)

• Unlike formal languages, natural languages allow a great deal of flexibility

• Example:

  ...*is estimated at 2.5 million residents...*

  ...*current population of Rome is 2817000...*

  ...*Rome housed over 1 million inhabitants...*
Why is QA hard? (3/3)

• Answers could be spread across different documents

• Examples:
  – *Which European countries produce wine?*
    [Document A contains information about Italy, and document B about France]
  
  – *What does Bill Clinton’s wife do for a living?*
    [Document A explains that Bill Clinton’s wife is Hillary Clinton, and Document B tells us that she’s a politician]
Architecture of a QA system

Question Analysis

IR

documents/passages

Answer Extraction

corpus

question

query

answer-type

representation

answers

question

representation

passage

representation
Question Analysis

• **Input:**
  Natural Language Question

• **Output:**
  Expected Answer Type
  (Formal) Representation of Question

• **Techniques used:**
  Machine learning (classification), parsing
Document Analysis

• **Input:**
  Documents or Passages

• **Output:**
  (Formal) Representation of Passages that might contain the answer

• **Techniques used:**
  Tokenisation, Named Entity Recognition, Parsing
Answer Retrieval

• **Input:**
  Expected Answer Type
  Question (formal representation)
  Passages (formal representation)

• **Output:**
  Ranked list of answers

• **Techniques used:**
  Matching, Re-ranking, Validation
Example Run

1. **Question Analysis**
   - Receives the question from the user.
   - Generates the query for the IR system.
   - Requires the question representation.
   - Outputs the answer-type.
   - Outputs the question representation.

2. **IR**
   - Receives the query from the Question Analysis module.
   - Processes the query to find relevant documents or passages in the corpus.
   - Outputs the documents or passages.

3. **Document Analysis**
   - Receives the documents or passages from the IR system.
   - Analyzes the passages to extract relevant answers.
   - Requires the passage representation.
   - Outputs the answers.

4. **Answer Extraction**
   - Receives the answers from the Document Analysis module.
   - Processes the answers to format them for display.
   - Outputs the final answers to the user.
Example Run

How long is the river Thames?

Question Analysis

Query

IR

documents/passages

Document Analysis

answer-type

question representation

Answer Extraction

answers
Example Run

length river Thames

Question Analysis

Answer Extraction

Document Analysis

IR

corpus

documents/passages

query

question

answer-type

question representation

passage representation

answers
Example Run

- Question Analysis
  - question
  - answer-type
  - question representation

- Answer Extraction
  - answers

- Document Analysis
  - documents/passages
  - passage representation

- MEASURE
  - corpus
Example Run

Question Analysis

Answer(x) & length(y,x) & river(y) & named(y,thames)

Document Analysis

Answer(x) & length(y,x) & river(y) & named(y,thames)

Answer Extraction

query

question

answer-type

question representation

answers

corpus

passage representation
Example Run

A: NYT199802-31
B: APW199805-12
C: NYT200011-07
Example Run

A: 30(u) & mile(u) & length(v,u) & river(y)

B: 60(z) & centimeter(z) & height(v,z) & dog(z)

C: 230(u) & kilometer(u) & length(x,u) & river(x)
Example Run

Question Analysis

IR

Document Analysis

Answer Extraction

C: 230 kilometer
A: 30 miles
B: 60 centimeter

corpus

documents/passages

passage representation

query

answes

question

C: 230 kilometer
A: 30 miles
B: 60 centimeter
Evaluating QA systems

• International evaluation campaigns for QA systems (open domain QA):
  – TREC (Text Retrieval Conference)
    http://trec.nist.gov/
  – CLEF (Cross Language Evaluation Forum)
    http://clef-qa.itc.it/
  – NTCIR (NII Test Collection for IR Systems)
    http://www.slt.atr.jp/CLQA/
TREC-type questions

• Factoid questions
  – Where is the Taj Mahal?

• List questions
  – What actors have played Tevye in `Fiddler on the Roof'?

• Definition/biographical questions
  – What is a golden parachute?
  – Who is Vlad the Impaler?
What is a correct answer?

• Example Factoid Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.
What is a correct answer?

• Example Factoid Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.
What is a correct answer?

• Example Factoid Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.

Inexact (under-informative)
What is a correct answer?

• Example Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.

Inexact (over-informative)
What is a correct answer?

• Example Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.
What is a correct answer?

• Example Question
  – When did Franz Kafka die?

• Possible Answers:
  – Kafka died in 1923.
  – Kafka died in 1924.
  – Kafka died on June 3, 1924 from complications related to Tuberculosis.
  – Ernest Watz was born June 3, 1924.
  – Kafka died on June 3, 1924.

Correct
Answer Accuracy

# correct answers

Answer Accuracy = ---------------------------------

# questions
Correct answers to list questions

Example List Question

*Which European countries produce wine?*

<table>
<thead>
<tr>
<th>System A:</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
</tr>
<tr>
<td>Italy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System B:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scotland</td>
</tr>
<tr>
<td>France</td>
</tr>
<tr>
<td>Germany</td>
</tr>
<tr>
<td>Italy</td>
</tr>
<tr>
<td>Spain</td>
</tr>
<tr>
<td>Iceland</td>
</tr>
<tr>
<td>Greece</td>
</tr>
<tr>
<td>the Netherlands</td>
</tr>
<tr>
<td>Japan</td>
</tr>
<tr>
<td>Turkey</td>
</tr>
<tr>
<td>Estonia</td>
</tr>
</tbody>
</table>
Evaluation metrics for list questions

• Precision (P):
  \[
  P = \frac{\text{# answers judged correct & distinct}}{\text{# answers returned}}
  \]

• Recall (R):
  \[
  R = \frac{\text{# answers judged correct & distinct}}{\text{# total answers}}
  \]

• F-Score (F):
  \[
  F = \frac{2 \times P \times R}{P + R}
  \]
Correct answers to list questions

Example List Question

Which European countries produce wine?

System A:

France
Italy

P = 1.00
R = 0.25
F = 0.40

System B:

Scotland
Germany
Spain
Greece
the Netherlands
Japan
Turkey
Estonia
France
Italy
Iceland

P = 0.64
R = 0.88
F = 0.74
### Other evaluation metrics

**System A: Ranked answers (Accuracy = 0.2)**

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>….</th>
<th>Qn</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A2</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A3</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A4</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A5</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
</tbody>
</table>

**System B: Ranked answers (Accuracy = 0.1)**

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>….</th>
<th>Qn</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A2</td>
<td>C</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>W</td>
<td>….</td>
</tr>
<tr>
<td>A3</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A4</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
<tr>
<td>A5</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>….</td>
<td>W</td>
</tr>
</tbody>
</table>
Mean Reciprocal Rank (MRR)

• Score for an individual question:
  – The reciprocal of the rank at which the first correct answer is returned
  – 0 if no correct response is returned

• The score for a run:
  – Mean over the set of questions in the test
# MRR in action

System A: \( \text{MRR} = \frac{0.2 + 1 + 1 + 0.2}{10} = 0.24 \)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A2</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A3</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A4</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A5</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
</tbody>
</table>

System B: \( \text{MRR} = \frac{0.5 + 0.33 + 0.5 + 0.25 + 1 + 0.5 + 0.5 + 0.5}{10} = 0.42 \)

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A2</td>
<td>C</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
<td>C</td>
</tr>
<tr>
<td>A3</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A4</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>C</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
<tr>
<td>A5</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td>W</td>
<td></td>
<td>W</td>
</tr>
</tbody>
</table>
Open-Domain Question Answering

• TREC QA Track
  – Factoid questions
  – List questions
  – Definition questions

• State-of-the-Art
  – Hard problem
  – Only few systems with good results
Architecture of a QA system

- Question Analysis
  - query
  - answer-type
  - passage representation
- IR
  - documents/passages
- Document Analysis
  - passage representation
- Answer Extraction
  - answers

- question
  - corpus
Architecture of a QA system

Question Analysis

IR

Document Analysis

Answer Extraction

corpus

documents/passages

query

question

answer-type

question representation

passage representation

answers
QUESTION ANALYSIS

• Question Classification; NLP techniques for question analysis; Tokenisation; Lemmatisation; POS-tagging; Parsing; Question Expansion (through wordnet).
Question TYPE (example)

• How many islands does Italy have?
• When did Inter win the Scudetto?
• What are the colours of the Lithuanian flag?
• Where is St. Andrews located?
• Why does oil float in water?
• How did Frank Zappa die?
• Name the Baltic countries.
• Which seabird was declared extinct in the 1840s?
• Who is Noam Chomsky?
• List names of Russian composers.
• Edison is the inventor of what?
• How far is the moon from the sun?
• What is the distance from New York to Boston?
• How many planets are there?
• What is the exchange rate of the Euro to the Dollar?
• What does SPQR stand for?
• What is the nickname of Totti?
• What does the Scottish word “bonnie” mean?
• Who wrote the song “Paranoid Android”? 
In how many categories would you classify the previous questions?
Distinguishing Questions Syntactically

• Wh-questions:
  – Where was Franz Kafka born?
  – How many countries are member of OPEC?
  – Who is Thom Yorke?
  – Why did David Koresh ask the FBI for a word processor?
  – How did Frank Zappa die?
  – Which boxer beat Muhammed Ali?
Syntactically Distinguished Questions

• Yes-no questions:
  – Does light have weight?
  – Scotland is part of England – true or false?

• Choice-questions:
  – Did Italy or Germany win the world cup in 1982?
  – Who is Harry Potter’s best friend – Ron, Hermione or Sirius?
Syntactically Distinguished Questions

• Imperative:
  – *Name four European countries that produce wine.*
  – *Give the date of birth of Franz Kafka.*

• Declarative:
  – *I would like to know when Jim Morrison was born.*
Semantically Distinguished Questions

• Divide questions according to their expected answer type
• Simple Answer-Type Typology:

PERSON (WHO?)
NUMERAL (HOW MANY?)
DATE (WHEN?)
MEASURE (HOW LONG..? WHAT IS THE HEIGHT..?)
LOCATION (WHERE?)
ORGANISATION (WHO?)
ENTITY (WHICH?)
Expected Answer Types

• DATE:
  – When was JFK killed?
  – In what year did Rome become the capital of Italy?
Expected Answer Types

• DATE:
  – *When was JFK killed?*
  – *In what year did Rome become the capital of Italy?*

• PERSON:
  – *Who won the Nobel prize for Peace?*
  – *Which rock singer wrote Lithium?*
Expected Answer Types

• DATE:
  – When was JFK killed?
  – In what year did Rome become the capital of Italy?

• PERSON:
  – Who won the Nobel prize for Peace?
  – Which rock singer wrote Lithium?

• NUMERAL:
  – How many inhabitants does Rome have?
  – What’s the population of Scotland?
Architecture of a QA system

- Question Analysis
  - question
  - answer-type
  - question representation
  - answers

- IR
  - query
  - documents/passages

- Answer Extraction
  - passage representation

- Document Analysis
Generating Query Terms

• Example 1:
  – Question: *Who discovered prions?*

• Query terms?

Text A: *Dr. Stanley Prusiner received the Nobel prize for the discovery of prions.*

Text B: *Prions are a kind of proteins that...*

A: is answer term is quite far
B: there is no answer term
Generating Query Terms

• Example 2:
  – Question: *When did Franz Kafka die?*

• Query terms?

• Text A: *Kafka died in 1924.*
  
  Text B: *Dr. Franz died in 1971*

  Partial matching (Franz kafka vrs Franz) maigh cause fatal errors
Generating Query Terms

• Example 3:
  • Question: *How did actor James Dean die?*

• Query terms?

• Answers
  – Text:
    
    *James Dean was killed in a car accident.*

    synonyms
Question Answering is difficult

- Needs morphologic, syntactic and semantic analysis
The Panda
A panda...

A panda walks into a cafe.
He orders a sandwich, eats it, then draws a gun and fires two shots in the air.
A panda...

“Why?” asks the confused waiter, as the panda makes towards the exit.
The panda produces a dictionary and tosses it over his shoulder.
“I am a panda,” he says. “Look it up.”
The panda’s dictionary

**Panda.** Large black-and-white bear-like mammal, native to China.
Eats shoots and leaves.
Eats, shoots and leaves.

VBZ  VBZ  CC  VBZ
Eats shoots and leaves.

VBZ  NNS  CC  NNS
Architecture of a QA system

Question Analysis

IR

Document Analysis

Answer Extraction

question

query

question representation

documents/passages

documents/passages

passage representation

answers

answer-type

question

answers

corpus
Natural Language is messy!

• EVERYTHING MATTERS
  – Punctuation
  – The way words are composed
  – The relationship between wordforms
  – The relationship between words
  – The structure of phrases

• This is where NLP (Natural Language Processing) comes in!
NLP Techniques (all needed in QA!)

- Tokenisation
- Lemmatisation
- Part of Speech Tagging
- Syntactic analysis (parsing)
- Semantic expansion
Tokenisation

• Tokenisation is the task that splits words from punctuation
  – Semicolons, colons  ; :
  – exclamation marks, question marks  ! ?
  – commas and full stops  ,
  – quotes “ ’ `

• Tokens are normally split by spaces
Tokenisation: combined words

• Combined words are split
  – I’d → I’d
  – country’s → country’s
  – won’t → will n’t
  – “don’t!” → “do n’t ! “

• Some Italian examples
  – gliel’ha detto → gli lo ha detto
  – posso prenderlo → posso prendere lo
Difficulties with tokenisation

• Abbreviations, acronyms
  – When was the U.S. invasion of Haiti?

• In particular if the abbreviation or acronym is the last word of a sentence
  – Look at next word: if in uppercase, then assume it is end of sentence
  – But think of cases such as Mr. Jones
Why is tokenisation important?

• To look up a word in an electronic dictionary (such as WordNet)

• For all subsequent stages of processing
  – Lemmatisation
  – Parsing
NLP Techniques

- Tokenisation
- Lemmatisation
- Part of Speech Tagging
- Syntactic analysis (parsing)
- WordNet
Lemmatisation

• Lemmatising means
  – grouping morphological variants of words under a single headword

• For example, you could take the words
  am,
  was,
  are,
  is,
  were, and
  been together under the word be
Lemmatisation

- Lemmatising means
  - grouping morphological variants of words under a single headword

- For example, you could take the words

  am, was, are, is, were, and been together under the word be
Lemmatisation

- Using linguistic terminology, the variants taken together form the **lemma of a lexeme**
- Lexeme: a “lexical unit”, an abstraction over specific constructions
- Other examples:
  
  dying, die, died, dies → die
  car, cars → car
  man, men → man
NLP Techniques

• Tokenisation
• Lemmatisation
• Part of Speech Tagging
• Syntactic analysis (parsing)
• Semantic analysis
Traditional parts of speech

- Verb
- Noun
- Pronoun
- Adjective
- Adverb
- Preposition
- Conjunction
- Interjection
Parts of speech in NLP

CLAWS1 (132 tags)

Examples:

NN singular common noun (boy, pencil ...

NNS$ genitive singular common noun (boy's, parliament's ...

NNP singular common noun with word initial capital (Austrian, American, Sioux, Eskimo ...

NNP$ genitive singular common noun with word initial capital (Sioux', Eskimo's, Austrian's, American's, ...

NNPS plural common noun with word initial capital (Americans, ...

NNPS$ genitive plural common noun with word initial capital (Americans‘, ...

NNS plural common noun (pencils, skeletons, days, weeks ...

NNS$ genitive plural common noun (boys', weeks' ...

NNU abbreviated unit of measurement unmarked for number (in, cc, kg ...

Penn Treebank (45 tags)

Examples:

JJ adjective (green, ...

JJR adjective, comparative (greener,...)

JJS adjective, superlative (greenest, ...

MD modal (could, will, ...

NN noun, singular or mass (table, ...

NNS noun plural (tables, ...

NNP proper noun, singular (John, ...

NNPS proper noun, plural (Vikings, ...

PDT predeterminer (both the boys)

POS possessive ending (friend's)

PRP personal pronoun (I, he, it, ...

PRPS possessive pronoun (my, his, ...

RB adverb (however, usually, naturally, here, good, ...

RBR adverb, comparative (better, ...)
What year did "Snow White" come out?
What WP
year NN
did VBD
" "
Snow NNP
White NNP
" "
come VB
out IN
? .
Why is POS-tagging important?

• To disambiguate words

• For instance, to distinguish “book” used as a noun from “book” used as a verb
  – I like that book
  – Did you book a room?

• Prerequisite for further processing stages, such as parsing
NLP Techniques

- Tokenisation
- Lemmatisation
- Part of Speech Tagging
- Syntactic analysis (parsing)
- Semantic analysis
What is Parsing

• **Parsing** is the process of assigning a syntactic structure to a sequence of words

• The syntactic structure is defined using a **grammar**

• A grammar contains of a set of symbols (terminal and non-terminal symbols) and production rules (**grammar rules**)

• The **lexicon** is built over the terminal symbols (i.e., the words)
Syntactic Categories

• The non-terminal symbols correspond to syntactic categories
  – Det (determiner)
  – N (noun)
  – IV (intransitive verb)
  – TV (transitive verb)
  – PN (proper name)
  – Prep (preposition)
  – NP (noun phrase)  
    – the car
  – PP (prepositional phrase)  
    – at the table
  – VP (verb phrase)  
    – saw a car
  – S (sentence)  
    – Mia likes Vincent
**Example Grammar**

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Grammar Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det: <em>which, a, the,</em>...</td>
<td>S $\rightarrow$ NP VP</td>
</tr>
<tr>
<td>N: <em>rock, singer,</em>...</td>
<td>NP $\rightarrow$ Det N</td>
</tr>
<tr>
<td>IV: <em>die, walk,</em>...</td>
<td>NP $\rightarrow$ PN</td>
</tr>
<tr>
<td>TV: <em>kill, write,</em>...</td>
<td>N $\rightarrow$ N N</td>
</tr>
<tr>
<td>PN: <em>John, Lithium,</em>...</td>
<td>N $\rightarrow$ N PP</td>
</tr>
<tr>
<td>Prep: <em>on, from, to,</em>...</td>
<td>VP $\rightarrow$ TV NP</td>
</tr>
<tr>
<td></td>
<td>VP $\rightarrow$ IV</td>
</tr>
<tr>
<td></td>
<td>PP $\rightarrow$ Prep NP</td>
</tr>
<tr>
<td></td>
<td>VP $\rightarrow$ VP PP</td>
</tr>
</tbody>
</table>
The Parser

• A **parser** automates the process of parsing
• The input of the parser is a string of words (possibly annotated with POS-tags)
• The output of a parser is a **parse tree**, connecting all the words
• The way a parse tree is constructed is also called a **derivation**
Derivation Example

Which rock singer wrote Lithium
Lexical stage

Which rock singer wrote Lithium
Use rule: $\text{NP} \rightarrow \text{Det N}$
Use rule: NP $\rightarrow$ PN

```
  NP
  /   \
Det N   N
  |   TV|
Which rock singer wrote Lithium
```
Use rule: VP $\rightarrow$ TV NP

Which rock singer wrote Lithium
Backtracking

Which rock singer wrote Lithium
Use rule: $N \rightarrow N \ N \ N$

Which rock singer wrote Lithium
Use rule: NP → Det N

```
NP
  |  
  |  
  |  
  |  
NP
  |  
  |  
  |  
  |  
VP
  |  
  |  
  |  
  |  
NP
  |  
  |  
  |  
  |  

Which  rock  singer  wrote  Lithium
```
Use rule $S \rightarrow NP \ VP$

Which *rock* singer *wrote* *Lithium*
Which rock singer wrote Lithium
Using a parser

- Normally expects tokenised and POS-tagged input

- Example of wide-coverage parsers:
  - Charniak parser
  - Stanford parser
  - Collins parser
  - RASP (Carroll & Briscoe)
  - CCG parser (Clark & Curran)
Stanford parser

• [http://nlp.stanford.edu/software/lex-parser.shtml](http://nlp.stanford.edu/software/lex-parser.shtml) (requires java 5)
Stanford Parser

Please enter a sentence to be parsed:
get back from the kitchen to the front door

Your query

get back from the kitchen to the front door

Tagging

get/VB back/RP from/IN the/DT kitchen/NN to/TO the/DT front/JJ door/NN

Parse

(ROOT
  (S
   (VP (VB get)
    (PRT (RP back))
    (PP (IN from)
     (NP
      (NP (DT the) (NN kitchen))
      (PP (TO to)
       (NP (DT the) (JJ front) (NN door))))))))

Typed dependencies

prt(get-1, back-2)
prep(get-1, from-3)
det(kitchen-5, the-4)
pobj(from-3, kitchen-5)
prep(kitchen-5, to-6)
det(door-9, the-7)
amod(door-9, front-8)
pobj(to-6, door-9)

Typed dependencies, collapsed

prt(get-1, back-2)
det(kitchen-5, the-4)
prep_from(get-1, kitchen-5)
det(door-9, the-7)
amod(door-9, front-8)
pobj_to(kitchen-5, door-9)

Statistics

Tokens: 9
Time: 0.060 s
NLP Techniques

• Tokenisation
• Lemmatisation
• Part of Speech Tagging
• Syntactic analysis (parsing)
• Semantics (WordNet, Framenet, Verbnet..)
Semantics

• Word sense disambiguation (plant living organism vrs plant building)
• Synonym expansion (Rome, Roma, Eternal City, Italian Capital, capital of Italy)
• Hypernym expansion (tylenol \(\rightarrow\) analgesic)
• Semantic parsing
Sources of semantic knowledge

- WordNet
- Framenet
- VerbNet
- Wikipedia (mostly unstructured, extremely high coverage of human knowledge!!)
WORDnet


Word to search for: plant  Search WordNet

Display Options: (Select option to change)  Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) plant, works, industrial plant (buildings for carrying on industrial labor) "they built a large plant to manufacture automobiles"
  - direct hyponym / full hyponym
  - domain term category
  - direct hypernym / inherited hypernym / sister term
    - S: (n) building complex, complex (a whole structure (as a building) made up of interconnected or related structures)
    - S: (n) structure, construction (a thing constructed; a complex entity constructed of many parts) "the structure consisted of a series of arches"; "she worked on the construction"
    - S: (n) artifact, artefact (a man-made object taken as a whole)
    - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a whole"
    - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
      - S: (n) physical entity (an entity that has physical existence)
    - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
- S: (n) plant, flora, plant life (botany) a living organism lacking the power of locomotion
- S: (n) plant (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) plant (something planted secretly for discovery by another) "he police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"

Verb
Motion_directional

Definition:

In this frame a Theme moves in a certain Direction which is often determined by gravity or other means. The paper FELL to the floor. The girl DROPPED 13 stories to her death.

FEs:

Core:

Area [Area]

This FE identifies the general Area in which motion takes place within. Soot-contaminated snow FALLS in Slovakia's mountains.

Direction [dir]

The direction of motion of the Theme Direction characterizes the movement or path of the Theme.

FRAMES

FRAMES description
Examples using WordNet in question type classification

• Which rock singer ...
  – singer is a hyponym of person, therefore expected answer type is PERSON

• What is the population of ...
  – population is a hyponym of number, hence answer type NUMERAL
Which rock singer wrote Lithium

Input: parse trees for the query
Identify question type

PERSON? rock singer wrote Lithium
QLF (quasi-logical forms)

1. Map the tree into a graph (e.g. Stanford parser’s dg)
QLF

2. Convert each node into an unary atom

\[ n_1, n_2, n_3, n_4, x \] are Skolem constants
3. Convert each edge into a binary atom

QLF: \( \text{PERSON}(x) \land \text{type}(x,n2) \land \text{mod}(n2,n4) \land \text{rock}(n4) \land \text{singer}(n2) \land \text{subj}(n1,n2) \land \text{write}(n1) \land \text{obj}(n1,n3) \land \text{Lithium}(n3) \)
QLF

Alternatively: verbs and prepositions are converted into binary or n-ary predicates, nouns in unary predicates

QLF: PERSON(e1) & rock(e1) & singer(e1) & write (e1,e2) & Lithium(e2)
Summary so far

• Classify question (e.g. PERSON)
• Extract search terms from question, possibly with sense expansion (e.g. rock singer write Lithium)
• Transform into QLF (as before)
• Search trough an IR (or Search Engine) matching documents, using query search terms
Architecture of a QA system

1. **Question Analysis**
   - Input: Question
   - Output: Answer-type, Question representation

2. **Document Analysis**
   - Input: Documents/Pages
   - Output: Passage representation

3. **IR**
   - Input: Query
   - Output: Documents/Pages

4. **Answer Extraction**
   - Input: Answers
   - Output: Answers

5. **Corpus**
   - Input: Collection of documents
Architecture of a QA system
Document analysis

- Named entity recognition
- Anaphora resolution
- Selecting the right “passage”
- Semantic Analysis (this is the same as for question)
Recall the Answer-Type Taxonomy

• We divided questions according to their expected answer type
• Simple Answer-Type Typology

PERSON
NUMERAL
DATE
MEASURE
LOCATION
ORGANISATION
ENTITY
Matching answer type=Named Entity Recognition

• In order to make use of the answer types, we need to be able to **recognise named entities** of the same types in the corpus

<table>
<thead>
<tr>
<th>PERSON</th>
<th>NUMERAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE</td>
<td>MEASURE</td>
</tr>
<tr>
<td>LOCATION</td>
<td>ORGANISATION</td>
</tr>
<tr>
<td>ENTITY</td>
<td></td>
</tr>
</tbody>
</table>
Italy’s business world was rocked by the announcement last Thursday that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
Italy’s business world was rocked by the announcement that Mr. Verdi would leave his job as vice-president of Music Masters of Milan, Inc to become operations director of Arthur Andersen.
Anaphora resolution
What is anaphora?

• Relation between a pronoun and another element in the same or earlier sentence

• Anaphoric pronouns:
  – he, she, it, they

• Anaphoric noun phrases:
  – the country,
  – that idiot,
  – his hat, her dress
Anaphora (pronouns)

• Question: What is the biggest sector in Andorra’s economy?

• Corpus: Andorra is a tiny land-locked country in southwestern Europe, between France and Spain. Tourism, the largest sector of its tiny, well-to-do economy, accounts for roughly 80% of the GDP.

• Answer: ?
Anaphora (definite descriptions)

• Question:
  What is the biggest sector in Andorra’s economy?

• Corpus:
  Andorra is a tiny land-locked country in southwestern Europe, between France and Spain. Tourism, the largest sector of the country’s tiny, well-to-do economy, accounts for roughly 80% of the GDP.

• Answer: ?
Anaphora Resolution

- Anaphora Resolution is the task of finding the **antecedents** of anaphoric expressions
- Example system:
  - http://clg.wlv.ac.uk/MARS/
Anaphora (pronouns)

• Question: 
  What is the biggest sector in Andorra’s economy?

• Corpus: 
  Andorra is a tiny land-locked country in southwestern Europe, between France and Spain. Tourism, the largest sector of Andorra’s tiny, well-to-do economy, accounts for roughly 80% of the GDP.

• Answer: Tourism
YOUR TURN

• Think of a simple heuristic for anaphora resolution
Architecture of a QA system
Answer Extraction

- Query/passage matching
- Reranking
- Sanity checking
Matching

• Given a question and an expression with a potential answer, calculate a matching score
  \[ S = \text{match}(Q,A) \]
  that indicates how well Q matches A

• Example
  – Q: When was Franz Kafka born?
  – A_1: Franz Kafka died in 1924.
  – A_2: Kafka was born in 1883.
Semantic Matching

Q:
TIME(X)
franz(Y)
kafka(Y)
born(E)
subj(E,Y)

A₁:
franz(x1)
kafka(x1)
die(x3)
subj(x3,x1)
in(x3,x2)
1924(x2)

Can be seen as an UNIFICATION process
Semantic Matching

Q: TIME(X)
   franz(Y)
   kafka(Y)
   born(E)
   subj(E,Y)

A₁: franz(x₁)
    kafka(x₁)
    die(x₃)
    subj(x₃,x₁)
    in(x₃,x₂)
    1924(x₂)

X=x₂
Semantic Matching

Q:

TIME(x2)
franz(Y)
kafka(Y)
born(E)
subj(E,Y)

A₁:
franz(x1)
kafka(x1)
die(x3)
subj(x3,x1)
in(x3,x2)
1924(x2)

Y=x1
Q: TIME(x2) 
franz(Y) 
kafka(x1) 
born(E) 
subj(E,x1) 

A: 
franz(x1) 
kafka(x1) 
die(x3) 
agent(x3,x1) 
in(x3,x2) 
1924(x2) 

Y=x1
Semantic Matching

Q: TIME(x2)
    franz(x1)
    kafka(x1)
    born(E)
    subj(E,x1)

A₁: franz(x1)
    kafka(x1)
    die(x3)
    agent(x3,x1)
    in(x3,x2)
    1924(x2)

Match score = 3/5 = 0.60
Semantic Matching

Q:

TIME(X)
franz(Y)
kafka(Y)
born(E)
subj(E,Y)

A₂:
kafka(x1)
born(x3)
subj(x3,x1)
in(x3,x2)
1883(x2)
Semantic Matching

Q:

TIME(x2)
franz(Y)
kafka(Y)
born(E)
subj(E,Y)

A2:
kafka(x1)
born(x3)
subj(x3,x1)
in(x3,x2)
1883(x2)

X=x2
Semantic Matching

Q:
TIME(x2)
franz(Y)
kafka(x2)
born(E)
subj(E,Y)

A2:
kafka(x1)
born(x3)
patient(x3,x1)
in(x3,x2)
1883(x2)

Y=x1
Semantic Matching

Q: TIME(x2)
   franz(x1)
   kafka(x1)
   born(E)
   subj(E,x1)

A2: kafka(x1)
    born(x3)
    subj(x3,x1)
    in(x3,x2)
    1883(x2)

E=x3
Semantic Matching

Q:
TIME(x2)
franz(x1)
kafka(x1)
born(x3)
subj(x3,x1)

A_2:
kafka(x1)
born(x3)
subj(x3,x1)
in(x3,x2)
1883(x2)

E=x3
Semantic Matching

Q:
TIME(x2)
franz(x1)
kafka(x1)
born(x3)
subj(x3,x1)

A₂:
kafka(x1)
born(x3)
patient(x3,x1)
in(x3,x2)
1883(x2)

Match score = 4/5 = 0.8
Matching Techniques

• Weighted matching
  – Higher weight for named entities

• WordNet
  – Hyponyms

• Inferences rules
  – Example:
    \[
    \text{BORN}(E) \& \text{IN}(E,Y) \& \text{DATE}(Y) \rightarrow \text{BIRTHDAY}(E) \& \text{IN}(E,Y) \& \text{DATE}(Y)
    \]
Reranking
Reranking

• Most QA systems first produce a list of possible answers...
• This is usually followed by a process called **reranking**
• Reranking promotes correct answers to a higher rank
Factors in reranking

• Matching score
  – The better the match with the question, the more likely the answers

• Frequency
  – If the same answer occurs many times, it is likely to be correct
Sanity Checking

Answer should be informative

Q: Who is Tom Cruise married to?
A: Tom Cruise

Q: Where was Florence Nightingale born?
A: Florence
Answer Validation

• Given a ranked list of answers, some of these might not make sense at all
• Promote answers that make sense

• How?
• Use even a larger corpus!
  – “Sloppy” approach
  – “Strict” approach
The World Wide Web
Answer validation (sloppy)

• Given a question Q and a set of answers $A_1 \ldots A_n$
• For each i, generate query $Q A_i$
• Count the number of hits for each i
• Choose $A_i$ with most number of hits
• Use existing search engines
  – Google, AltaVista
  – Magnini et al. 2002 (CCP)
  – Btw: WATSON does this!!
Corrected Conditional Probability

• Treat Q and A as a bag of words
  – Q = content words question
  – A = answer

  \[
  \text{hits}(A \text{ NEAR } Q)
  \]

• \(\text{CCP}(Qsp,Asp) = \frac{\text{hits}(A) \times \text{hits}(Q)}{}

• Accept answers above a certain CCP threshold
Answer validation (strict)

• Given a question $Q$ and a set of answers $A_1...A_n$
• Create a declarative sentence with the focus of the question replaced by $A_i$
• Use the strict search option in Google
  – High precision
  – Low recall
• Any terms of the target not in the sentence as added to the query
Example

• TREC 99.3
  Target: Woody Guthrie.
  Question: Where was Guthrie born?
• Top-5 Answers:
  1) Britain
  * 2) Okemah, Okla.
  3) Newport
  * 4) Oklahoma
  5) New York
Example: generate queries

• TREC 99.3
  Target: Woody Guthrie.
  Question: Where was Guthrie born?

• Generated queries:
  1) “Guthrie was born in Britain”
  2) “Guthrie was born in Okemah, Okla.”
  3) “Guthrie was born in Newport”
  4) “Guthrie was born in Oklahoma”
  5) “Guthrie was born in New York”
Example: add target words

• TREC 99.3
  Target: **Woody Guthrie.**
  Question: **Where was Guthrie born?**

• Generated queries:
  1) “Guthrie was born in Britain” Woody
  2) “Guthrie was born in Okemah, Okla.” Woody
  3) “Guthrie was born in Newport” Woody
  4) “Guthrie was born in Oklahoma” Woody
  5) “Guthrie was born in New York” Woody
Example: morphological variants

TREC 99.3

Target: Woody Guthrie.

Question: Where was Guthrie born?

Generated queries:

“Guthrie is OR was OR are OR were born in Britain” Woody
“Guthrie is OR was OR are OR were born in Okemah, Okla.” Woody
“Guthrie is OR was OR are OR were born in Newport” Woody
“Guthrie is OR was OR are OR were born in Oklahoma” Woody
“Guthrie is OR was OR are OR were born in New York” Woody
Example: google hits

TREC 99.3
Target: Woody Guthrie.
Question: Where was Guthrie born?

Generated queries:
“Guthrie is OR was OR are OR were born in Britain” Woody 0
“Guthrie is OR was OR are OR were born in Okemah, Okla.” Woody 10
“Guthrie is OR was OR are OR were born in Newport” Woody 0
“Guthrie is OR was OR are OR were born in Oklahoma” Woody 42
“Guthrie is OR was OR are OR were born in New York” Woody 2
Example: reranked answers

TREC 99.3
Target: Woody Guthrie.
Question: Where was Guthrie born?

Original answers
1) Britain
* 2) Okemah, Okla.
3) Newport
* 4) Oklahoma
5) New York

Reranked answers
* 4) Oklahoma
* 2) Okemah, Okla.
5) New York
1) Britain
3) Newport
Summary

• Introduction to QA
  – Typical Architecture, Evaluation
  – Types of Questions and Answers

• Use of general NLP techniques
  – Tokenisation, POS tagging, Parsing
  – NER, Anaphora Resolution

• QA Techniques
  – Matching
  – Reranking
  – Answer Validation
Where to go from here

• Producing answers in real-time
• Improve accuracy
• Answer explanation
• User modelling
• Speech interfaces
• Dialogue (interactive QA)
• Multi-lingual QA