Course 12
Evaluation metrics and Question-Answering systems

Slides offered by Adrian Iftene
Content

- NLP Systems Evaluation
- Information Retrieval
- Information Extraction
- Question Answering
  - Introduction
  - System components
    - Background knowledge indexing
    - Index creation and information retrieval
    - Answer extraction
  - Results
  - Error analysis
- Conclusions
NLP Systems need evaluation

- “An important recent development in NLP has been the use of much more rigorous standards for the evaluation of NLP systems”
  Manning and Schutze

- To be published, all research must:
  ◦ establish a **baseline**, and
  ◦ quantitatively show that it improves on the baseline
NLP Systems – ways of doing evaluation

“How well does the system work?”

Possible domains for evaluation
- Processing time of the system
- Space usage of the system
- Human satisfaction
- Correctness of results

Measures: (Accuracy, Error), (Precision, Recall, F-measure)
Comparing the output of the system with a gold standard we can verify what is correct.

The results of a system are marked as:
- **Correct**: matches the gold standard
- **Incorrect**: otherwise

\[
\text{accuracy} = \frac{\text{number correct}}{\text{number of results}}
\]

\[
\text{error} = \frac{\text{number wrong}}{\text{number of results}}
\]
Accuracy and Error – Example

“ مساء الخير” ➔ Arabic ✓
“Bon soir!” ➔ French ✓
“Good evening!” ➔ Spanish ×

- Accuracy = 66.66 %
- Error = 33.33 %
Precision and Recall

- Precision and Recall are set-based measures.
- They evaluate the quality of some set membership, based on a reference set membership.
- **Precision**: what proportion of the *retrieved documents* is relevant?

\[
\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
\]

- **Recall**: what proportion of the *relevant documents* is retrieved?

\[
\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}
\]
Precision and Recall – Example (1)

Precision = 4 / 10 = 40%
Recall = 4 / 14 = 28.57%
Precision and Recall – Example (2)

Precision = $\frac{14}{20} = 70\%$
Recall = $\frac{14}{14} = 100\%$
Precision and Recall – Example (3)

Precision = 0 / 6 = 0 %
Recall = 0 / 14 = 0 %
F-measure (F-score or F1-score)

- F-measure is a measure of a test's accuracy, and it considers both the precision \( p \) and the recall \( r \)

- General formula: \( F_{\beta} = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})} \)

- F1-measure: \( F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})} \).

- F2-measure = ?
Question Answering (QA): a QA system takes as input a question in natural language and produces one or more ranked answers from a collection of documents.
Question Answering – Modules

- QA systems normally adhere to the pipeline architecture composed of three main modules (Harabagiu and Moldovan, 2003):
  - **question analysis** – *the results are keywords, answer and question type, focus*
  - **paragraph retrieval** – *the results are a set of relevant candidate paragraphs/sentences from the document collection*
  - **answer extraction** – *the results are a set of candidate answers ranked using likelihood measures*
QA – Question Type

- Harabagiu and Moldovan, 2003:
  - **Factoid** – “Who discovered the oxygen?”, “When did Hawaii become a state?” or “What football team won the World Coup in 1992?”
  - **List** – “What countries export oil?” or “What are the regions preferred by the Americans for holidays?”.
  - **Definition** – “What is a quasar?” or “What is a question–answering system?”
  - **How, Why**, hypothetical, semantically constrained, polar (Yes/No) and cross–lingual questions
QA – Answer Type

- **Person** – "What", "Who", "Whom", "With who"
- **Location** (City, Country, and Region) – "What state/city“, "From where”, "Where“
- **Organization** – "Who produced“, "Who made“
- **Temporal** (Date and Year) – “When”
- **Measure** (Length, Surface and Other) – “How many/much”
- **Count** – "How many/much“
- **Yes/No** – “Did you fear that?”, “Are you blue?”
QA – Search collection

- Local collections, internal organization documents, newspapers, Internet

- **Closed-domain** – deals with questions from a specific domain (medical, baseball, etc.). Can exploit domain-specific knowledge (ontologies, rules, disambiguation)

- **Open-domain** – general question about anything. Can use general knowledge about the world
Examples of such systems

- The first QA systems have been created in the 60s:
  - **BASEBALL** (Green 1963) – answer questions about baseball games

- **LUNAR** (Woods, 1977) – geological analysis of rocks returned by the Apollo moon missions

- **IURES** (Cristea, Tufiș, Mihăiescu, 1985) – medical domain
QA – Systems of today

- Powerset: http://www.powerset.com/ (http://www.bing.com/)
- Assimov the chat bot: http://talkingrobot.org/b/
- NSIR: http://tangra.si.umich.edu/clair/NSIR/html/nsir.cgi
- START (The first question answering system): http://start.csail.mit.edu/
Hello Adrian,
My name is Asimov.

Hello Asimov.
How are you?
Hello Adrian.
I am fine, thank you.
I am always the same. I don't have feelings. I have no feelings or moods.

Submit
20

NSIR
Question Answering System

Sample Questions

1. What year did Wilt Chamberlain score 100 points?
2. How many chromosomes does a human zygote have?
3. Who is Tom Cruise married to?
4. What is South Africa’s total population?
5. What is Canada’s most populous city?
6. Which country exports the most tea?
7. What state has the most Indians?
8. George Bush purchased a small interest in which baseball team?
9. What county is Modesto, California in?

Type your question below
or choose from TREC-8  TREC-9  TREC-2001  TREC-2002

Who is Tom Cruise married to?

Search Engine / Index: help
No. of Hits: help
Answer Length: help
No. of Answers: help
Answer help (for more, use “,” to delimit multiple answers)
Cached results: help
Debug information: help

This is a PERSON type question

Ranked Answers (click to see context)

1) (What are those numbers?)

Top Web Documents Retrieved By Google

FARR1  FHS1  TRDR1  TDRTR  PREC1  MRR1  TRWR1  NTRWR1
Possible Answers

**How do I make French toast with only Eggs and a bit of milk and bread?**

...my first time making French toast can someone maybe put it step by step on how to make it and also I don't have Vanilla something... is (Sugar, Milk, Bread, and eggs, and butter) that's it so how do I make it like that? Please help...

**How many calories are on a piece of French toast?**

I'm doing a school project to see how many calories I eat a day and how much... this is the only thing I can't find! Its normal white bread dipped in eggs and a skim milk and cooked on a pan with a bit of butter... my mom makes it so I don't know...

**Did I give a good answer to this question??**

...with lots of butter Milk is really good but make sure its 2% or less...Lunch... make sure again its on wheat bread and try to add other... terrible, cut them out of your diet or at least don't have if its good Cereal and toast is good like I said above if its... and you don't load it with lots of butter What... you getting? Eggs are good 2 but how are you cooking... up at 5:45 to exercise... do you have no time to exercise... a week Pancakes, French Toast, and Waffles... have to have them only have a little bit Dinner Subway is good...

**More Resources from Wiki and Web**
START's reply

--- Who is Ion Ilescu?

**Ion Ilescu**

Ion Ilescu (born March 23, 1962) is a computer scientist and former member of the Central Intelligence Agency (CIA). He worked as a senior analyst for the CIA from 1989 to 2000. Ilescu was later employed by the Finnish Institute of Nuclear Technology (SCK-CEN). He is currently a professor at the University of Washington.

**Source:** Global Gazetteer

--- Where is Bucharest?

Bucharest, Romania is located at 232 feet above sea level.

**Source:** START KB

--- Geography

- Go back to the START dialog window

The coordinates of Bucharest, Romania are 44.43 N, 26.1 E.

**Source:** START KB
QA – Competitions

  - Coordination: Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy
  - Romanian Institute for Computer Science, Romania

  - National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA
An excerpt from the gold standard file
Our group participates at CLEF exercises since 2006:
- 2006 – Ro–En (English collection) – 9.47% right answers
- 2007 – Ro–Ro (Romanian Wikipedia) – 12 %
- 2008 – Ro–Ro (Romanian Wikipedia) – 31 %
- 2009 – Ro–Ro, En–En (JRC–Acquis) – 47.2 % (48.6%)
- 2010 – Ro–Ro, En–En, Fr–Fr (JRC–Acquis, Europarl) – 47.5% (42.5%, 27 %)
UAIC System components

Background knowledge

Lucene index 1

Questions processing:
- Lemmatization
- Stop words elimination
- NEs identification
- Lucene query

Identify relevant documents

Test data (documents, questions, possible answers)

Lucene indexes 2

Answers processing:
- Lemmatization
- Stop words elimination
- NEs identification
- Lucene query

Partial and global scores per answers

Test data (documents, questions, possible answers)
Background knowledge indexing

- The Romanian background knowledge has 161,279 documents in text format
  - 25,033 correspond to the AIDS topic
  - 51,130 to Climate Change topic
  - 85,116 to Music and Society topic

- The indexing component considers the name of the file and the text from it => Lucene index 1
Test data processing

- Test data was an XML file with 12 test documents
  - 4 documents for each of the three topics (12 in total)
  - 10 questions for each document (120 in total)
  - 5 possible answers for each question (600 in total)

- Test data processing involved 3 operations:
  - extracting documents
  - processing questions
  - processing possible answers
Test data processing – Extracting documents

- The content of $\langle doc \rangle \Rightarrow \langle topic id \rangle \langle reading test id \rangle \backslash 1..10$
Test data processing – Processing questions

- Stop words elimination
- Lemmatization
- Named Entity identification
- Lucene query building

```
<topic t id="1" t name="AIDS" r id="2" q id="3">
<string>Which African country did Bono Vox visit?</string>
<focus>country</focus>
<verb>visit</verb>
<noun></noun>
<adj></adj>
<nameEntities>African Bono Vox</nameEntities>
<query>African^3 country Bono^3 Vox^3 visit</query>
<questionType>FACTOID</questionType>
<answerType>OTHER</answerType>
```
Similar to processing questions +
We use ontology (Iftene and Balahur, 2008) for elimination of possible answers with low probability to be final answer (relation [is_located_in])

In which European cities has Annie Lennox performed?
We eliminate from the list of possible answers the answers with non-European cities (we replace non-European cities with the value XXXXX)

```xml
<answer a_id='1'>London^3 Paris^3</answer>
<answer a_id='2'>XXXX^3</answer>
<answer a_id='3'>XXXX^3</answer>
<answer a_id='4'>Edinburgh^3 Oslo^3</answer>
<answer a_id='5'>XXXX^3</answer>
```
Indexing relevant documents for questions

- For every question we index relevant documents returned by Lucene at the previous step and relevant documents saved from the initial test file.
Then in every index, we performed searches using Lucene queries associated to possible answers.

For every answer, we obtained a list of documents with Lucene relevance scores.

\(\text{Score2}(d, a)\) is the relevance score for document \(d\) when we search with the Lucene query associated to the answer \(a\).
Results 1

- Results of UAIC’s runs at question answering level

<table>
<thead>
<tr>
<th></th>
<th>Ro-Ro</th>
<th></th>
<th>En-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>answered right</td>
<td>30</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>answered wrong</td>
<td>85</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>total answered</td>
<td>115</td>
<td>30</td>
<td>62</td>
</tr>
<tr>
<td>unanswered right</td>
<td>0</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>unanswered wrong</td>
<td>0</td>
<td>66</td>
<td>42</td>
</tr>
<tr>
<td>unanswered empty</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>total unanswered</td>
<td>5</td>
<td>90</td>
<td>58</td>
</tr>
<tr>
<td><strong>Overall accuracy</strong></td>
<td>0.25</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td><strong>C@1 measure</strong></td>
<td>0.26</td>
<td>0.16</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Links

- Yes–no question: [http://en.wikipedia.org/wiki/Yes%E2%80%93no_question](http://en.wikipedia.org/wiki/Yes%E2%80%93no_question)
- Lecture 13: Evaluation: Precision and Recall
- Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology: