Is Question Answering an Acquired Skill?

presented by Arda Çelebi

07.28.2011
Questions and Answers

• Factoids
  – What is the capital of Australia?
  – Who is the president of Japan?
  – Where is Hawaii?
  – How many people live in China?
  – When did Apollo 11 land on the moon?

• Harder questions = not definite answers
  – Why do French people protest too much?
QA vs IR

• IR
  − Looking for set of relevant documents
  − Output is list of documents; representation is easy

• QA
  − Looking for the exact answer: you have one shot!
  − Output is one sentence in natural language; may require Natural Language Generation (NLG)
  − It may require reasoning
Q&A Systems

- AnswerBus; one of the oldest
- Watson by IBM at Jeopardy!
Structured Questions : SQL Query

```sql
select Name from People where BornIn='Izmir';
```

- What we are asking is well-defined
  - We are looking for a list of people
  - More specifically we are looking for people born in Izmir.
  - No other interpretation = no ambiguity
Natural language to Structured format

- **Basic Idea**: transform questions in natural language to a structured format (i.e. SQL)
- This structured format can be
  - Templates/Frames/Cases
  - Triple-based representation like Web Ontology Language (OWL)
Goals of the Paper

- Dividing Q&A task into two parts
  - Discovering features to identify answers
  - Learning to score potential answer snippets from a corpus and past question-answer example pairs
- Use off-the-shelf tools as much as possible
- Being able to reuse expertise from one corpus to a new corpus
Two sides of Q&A

- Understanding the question
  - converting question in NL to structured format

- Choosing the best possible answer
  - ranking the possible answers to get the best one
Selector & Atype

- Seeing the question as a SQL query
  - Selector = WHERE clause
  - Atype = COLUMN we are looking for

Example:

“who was born in Izmir?”  OR
“select Name from People where BornIn='Izmir'”

- **Selector** = Izmir
- **Atype** = Person/Name
What is an atype?

• minimal subclass of entities which will answer a given question

• two types of representations are considered
  – WordNet synsets
  – Syntactic surface patterns
    • eg. “at DD:DD”, “in DDDD”

• manifested through specific words (Where, When, Who) or by generic words (Which city)
How to select an atype?

**Method 1**: Use shallow parsing or chunking

```
(S (NP what American general)
 (VP is
  (VP buried
   (PP in
    (NP Salzburg)))))
```

- **Study**: Atype words are embedded in the noun phrases in the neighbourhood of either the aux or the main verb
How to select an atype? (Cont'd)

**Method 2**: Learning to map clues to atypes

- Unlike “which X” questions, when, where, who questions do not explicitly tell the atype
- Goal is to learn mappings between short token sequences in questions to atypes
What is a selector?

• Words that we expect will appear unchanged in a passage containing the answer
  – In case of SQL query “... where BornIn='Izmir'”, we expect that answer contains the word “Izmir”

• When submitting a query to search engines, we unconsciously determine selectors and include variety of non-selector alternatives
How to spot a selector?

- Depends on various features
  - Local features:
    - The part-of-speech (POS) of that word
    - The POS assigned to left and right neighbors of that word
    - Whether it starts with uppercase
  - Global features:
    - Whether it is a stop word
    - IDF value of that word
    - Number of senses that is has
    - Number of words that expresses the same sense
Learning to spot a selector

- Generate a feature vector for each word in the question
- Use different machine learning techniques (from WEKA data mining toolkit)
  - Decision trees (C4.5 and J48)
  - Logistic regression
Using selectors

• Two ways of using selectors
  1) to pad the initial keyword query
     • Candidate answers are retrieved by IR system Lucene
     • Returned documents are passages
  2) to re-rank the candidate passages
     • Considering the distance between answer zone and a selector
Passages

- **Passage**: a window of sentences from document(s) that might include desired answer

- **Answer zone**: portion of passage which is the actual answer
Scoring Passages

- IR system is used to retrieve passages that may include the answer
- Ranking of IR system not enough for QA
- Re-ranking = can be done by supervised learning or classification problem
  - Identifying each <question, passage> pair as positive or negative instance
  - Considering semantic similarity between answer zone of the passage and atype of question ... WordNet
Introducing the WordNet

- is lexical database for the English language
- groups words into sets of synonyms called *synsets*
- provides short and general definitions
- shows semantic relations between synsets
  - *isA* relation (hyponyms) : “canine is type of a dog”
  - *partOf* relation (holonyms) : “window is part of a building”
Online Access to WordNet @
http://wordnetweb.princeton.edu/perl/webwn

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for:  book  Search WordNet

Noun

- S: (n) book#1 (a written work or composition that has been published (printed on pages bound together)) "I am reading a good book on economics"
- S: (n) book#2, volume#3 (physical objects consisting of a number of pages bound together) "he used a large book as a doorstep"
- S: (n) record#5, record book#1, book#3 (a compilation of the known facts regarding something or someone) "Al Smith used to say, `Let's look at the record''; "his name is in all the record books"
- S: (n) script#1, book#4, playscript#1 (a written version of a play or other dramatic composition; used in preparing for a performance)
- S: (n) ledger#1, leger#1, account book#1, book of account#1, book#5 (a record in which commercial accounts are recorded) "they got a subpoena to examine our books"
WordNet IsA Hierarchy

- object
  - artifact
    - instrumentality
    - conveyance, transport
      - vehicle
        - wheeled vehicle
          - automotive, motor
            - car, auto, ...
          - truck
      - bike, bicycle
    - article
      - ware
        - table ware
          - cutlery, eating utensil
          - fork
Calculating similarity w/ WordNet

- Basic approach: get the shortest path between two and scale with longest path in the whole database (dissimilarity)

- Proposed approach in the paper
  - Get all the ancestor senses for compared pair
  - Use Jaccard overlap/similarity measure between two sets of senses
Experiments
Data Preparation

• Datasets from TREC QA track
  – Millions of docs and manually prepared <question,answer> pairs

• Sliding windows of 3 sentences selected as passages

• Passages are indexed by Lucene IR engine

• Questions and passages tokenized w/ GATE
Data Preparation (Cont'd)

- **GATE** also includes POS tagger which assigns one of 36 standard POS (eg. NN) to each token
- **Link Parser** is used for shallow parsing
- For learning tasks, **J48** decision tree and logistic regression tools used from **WEKA**
From question to atype

- Manual inspection of hundreds of queries showed that they extracted correct atype clues

<table>
<thead>
<tr>
<th>Question type</th>
<th>Total questions</th>
<th># correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>what</td>
<td>630</td>
<td>612</td>
</tr>
<tr>
<td>which</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>name</td>
<td>23</td>
<td>20</td>
</tr>
</tbody>
</table>

- Sample frequent connections from atype clues to WordNet synsets:
Spotting the selector

- J48 compared with Logistic regression

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>%correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>0.79</td>
<td>0.71</td>
<td>0.75</td>
<td>74.5</td>
</tr>
<tr>
<td>J48</td>
<td>0.84</td>
<td>0.78</td>
<td>0.81</td>
<td>80.5</td>
</tr>
</tbody>
</table>

- Decision tree showing intuitive rules to identify the selector in the question

```
POS0-0=adj
  POS0-1=noun
    NumSense0 <= 9
    NumLemma0 <= 2.5: selector
    NumLemma0 > 2.5: not-selector
    NumSense0 > 9: not-selector

POS0-0=verb
  NumLemma0 <= 1.82
  POS0-1=noun
  POS0+1=noun: selector
  POS0+1=noun: not-selector
  NumLemma0 > 1.82: not-selector
```
Passage Reranking

- Reranking is considered as classification of passages; whether they are positive or negative.
- By only using 2% - 4% of training data, they can settle near its best accuracy.
- Large number of negative instance makes recall low.

<table>
<thead>
<tr>
<th>TREC 2000</th>
<th>-1</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>51228</td>
<td>308</td>
</tr>
<tr>
<td>+1</td>
<td>5017</td>
<td>3359</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TREC 2002</th>
<th>-1</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>38396</td>
<td>51</td>
</tr>
<tr>
<td>+1</td>
<td>2169</td>
<td>550</td>
</tr>
</tbody>
</table>

- Precision ($P$) = $0.918 \pm 0.013$, $0.934 \pm 0.023$
- Recall ($R$) = $0.4 \pm 0.004$, $0.2 \pm 0.0038$
- $F_1$ = $0.56 \pm 0.002$, $0.33 \pm 0.004$
Passage Reranking (Cont'd)

Reranking pushes answers to the top ranks
Mean Reciprocal Rank

- Suppose $n_q \geq 1$ is the earliest rank of the passage at which the answer to question $q \in Q$ is found.
- Then MRR is $\frac{1}{|Q|} \sum_{q} \left( \frac{1}{n_q} \right)$.
- MRR is between 0 and 1 and a higher MRR is better.

<table>
<thead>
<tr>
<th>Passage ranked by</th>
<th>TREC 2000</th>
<th>TREC 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR score (Lucene)</td>
<td>0.377</td>
<td>0.249</td>
</tr>
<tr>
<td>LR score</td>
<td>0.71 ± 0.001</td>
<td>0.565 ± 0.001</td>
</tr>
</tbody>
</table>
How generalizable?

<table>
<thead>
<tr>
<th>Trained on</th>
<th>Tested on</th>
<th>Reranked MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>2002</td>
<td>0.565</td>
</tr>
<tr>
<td>2000</td>
<td>2002</td>
<td>0.539</td>
</tr>
<tr>
<td>2000+2001</td>
<td>2002</td>
<td>0.534</td>
</tr>
<tr>
<td>2000+2001+1999</td>
<td>2002</td>
<td>0.541</td>
</tr>
<tr>
<td>2000</td>
<td>2000</td>
<td>0.710</td>
</tr>
<tr>
<td>2002</td>
<td>2000</td>
<td>0.705</td>
</tr>
<tr>
<td>2002+2001</td>
<td>2000</td>
<td>0.627</td>
</tr>
<tr>
<td>2002+2001+1999</td>
<td>2000</td>
<td>0.693</td>
</tr>
</tbody>
</table>
Conclusion

- QA system that is built by wrapping a small and simple logic around text indexers, taggers, shallow parsers and classifiers
- It is not that building blocks are simple and free but the way they assemble them together is simple (only about 4,000 lines of JAVA code)
- Future Work
  - Better than single-shot selector model
  - Query expansion guided by QA-mining
Critique

• Pros
  – Simple approach; most of the system is off-the-shelf
  – Generalizable approach to new corpora
  – MRR is good metric for QA

• Cons
  – Title of the paper does not reflect the content
  – Not successful at comparing the performances with existing systems
  – Weak “Future Work” section
Thanks for listening :) 

Any questions?