A Data Driven Approach to Relevancy Recognition for Contextual Question Answering

Authors: Fan Yang (OGI)
Junlan Feng
Giuseppe Di Fabbrizio

Speaker: Junlan Feng
Outline

➢ *Motivations*
  • Previous research work
  • A data driven approach
  • Results
    – Results on TREC data
    – Results on HandQA data
  • Preliminary contextual information fusion
  • Future work
Motivations

• **WebTalk, a Research Project at AT&T Labs**
  
  – WebTalk is a system for analyzing unstructured information from company websites to support automatic creation of customer care dialog applications.

  – Question Answering is a key component technology.
    • Users often ask questions naturally as part of contextualized interaction.
    • Many questions that users frequently want answers for cannot be satisfied with a simple answer. By the nature, the question initiates a dialog.

• **Most available QA systems and QA technologies are limited to answer questions in isolation.**

• **Contextual question answering (QA), in which users’ information needs are satisfied through an interactive QA dialogue.**
Research Purpose

- To develop techniques for contextual QA
  - Relevancy recognition
    • Determine whether a question is relevant to the previous interaction context
  - Contextual information fusion:
    • Use contextual information to help retrieve answers
Outline

• Introduction
  ➢ Related research work
• A data driven approach
• Results
  – Results on TREC data
  – Results on HandQA data
• Preliminary contextual information fusion
• Future work
A Rule Based Algorithm


Given a sequence of questions $Q_1, \ldots, Q_i$

Syntactic Rules:

1. If $Q_i$ has a pronoun or possessive adjective, which has no references in the current question, $Q_i$ is a follow-up question.
2. If $Q_i$ has cue words such as “precisely” or “exactly”, $Q_i$ is a follow-up question.
3. If $Q_i$ does not contain any verbs, $Q_i$ is a follow-up question.

Semantic Rules:

4. Otherwise, calculate the semantic similarity measure of $Q_i$ as $\text{SimilarityMeasure}(Q_i) = \max f(j) \cdot \text{SentenceSimilarity}(Q_i; Q_{i-j})$
   Here $f(j)$ is a decay function. If the similarity measure is higher than a certain threshold, $Q_i$ is a follow-up question.
5. Otherwise, if answer is available, calculate the semantic similarity between $Q_i$ and the immediate previous answer $A_{i-1}$: $\text{SentenceSimilarity}(Q_i; A_{i-1})$. If it is higher than a certain threshold, $Q_i$ is a follow-up question that is related to the previous answer.

6. Otherwise, $Q_i$ begins a new topic.
Outline

• Introduction
• Previous research work
  ➢ A data driven approach
• Results
  – Results on TREC data
  – Results on HandQA data
• Preliminary contextual information fusion
• Future work
Feature Extraction

• For each question, we extract
  – Syntactic features (binary)
    • Pronoun: (exception “I”, “our”, “yours” …)
    • Noun:
    • Proper Noun:
    • Verb:
  – Semantic Similarity between Q and the context
    – Word similarity measures:
      • PATH: noun and verb
      • WUP: noun and verb [Wu & Palmer 1994]
      • LIN: noun and verb [Lin 1998]
      • VECTOR: noun, verb, and adjective [Patwardhan 2003]
Semantic Similarity Between Questions

• Questions:
  
  Current question: \( Q = \{w_1, w_2, \ldots, w_n\} \)
  
  A previous question: \( Q' = \{w'_1, w'_2, \ldots, w'_m\} \)

• Sentence-sentence similarity

\[
SentenceSimilarity(Q, Q') = \frac{1}{n} \sum_{1 \leq j \leq n} \left( \max_{1 \leq i \leq m} WordSimilarity(w_j, w'_i) \right)
\]

• Word-word similarity: based on Wordnet
  – PATH, LIN, WUP, VECTOR
Question Similarity Measurement

Given a sequence of questions \( Q_1, \ldots, Q_{ni} \)

\[
\text{Similarity}(Q_i, \text{Context}) = \max_{0 < j < i} (d(j) \times SS(Q_i, Q_{i-j}))
\]

\[
d(x) = 1 - \frac{1}{1 + e^{(n-x)}}
\]
is a decay function
Learning Algorithms

- A binary classification problem
  - Decision Tree (DT)
  - Adaboost
Outline

• Introduction
• Previous research work
• A data driven approach
  ➢ *Results on Relevancy Recognition*
    – Results on TREC data
    – Results on HandQA data
• Preliminary contextual information fusion
• Future work
TREC Data

• Training set
  – TREC 2004
  – 286 questions
  – 64 series

• Testing set
  – TREC 2001 context track
  – 42 questions
  – 10 series
# Results on TREC Data Using DT

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicted Class</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96.9%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.5%</td>
<td>97%</td>
</tr>
<tr>
<td>Follow</td>
<td></td>
<td>99.0%</td>
<td>56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>93%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Orange: performances using the rule-based algorithm
Tree

```
root
  ┌───────► pronoun = 0
  │         ┌───────► PATH noun < 0.31
  │         │         ┌───────► PATH noun > 0.31
  │         │         │         ┌───────► following-up
  │         │         │         │         ┌───────►
  │         │         │         │         │         ┌───────►
  │         │         │         │         │         │         ┌───────►
  │         │         │         │         │         │         │         ┌───────►
```

Error Analysis

- 2 failures to recognize follow-up question
  - Lack of semantic relations in WordNet

- 1 failure to recognize the first question
  - Over-fitting of decision tree learning
  - Adaboost?
## Results Using Adaboost

<table>
<thead>
<tr>
<th>True Class</th>
<th>Predicted Class</th>
<th>Training data</th>
<th>Testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First</td>
<td>Follow</td>
</tr>
<tr>
<td>First</td>
<td>First</td>
<td>64</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Follow</td>
<td>1</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td>First</td>
<td>65</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>Follow</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td>98.5%</td>
<td>99%</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td>99.0%</td>
<td></td>
</tr>
</tbody>
</table>
HandQA Data

• Characteristics
  – Real data collected from a customer care QA system
  – Repeat or rephrase questions. Examples:
    *How to make number non published*
    *Non published numbers*
    *How to make number non listed*
  – Noisy: typos, bad grammars, keywords, ...
  – 5908 questions
  – 2184 series
  – 90% data used for training, 10% for testing
Results on HandQA Data

<table>
<thead>
<tr>
<th>Class</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>75%</td>
<td>68%</td>
<td>73%</td>
<td>62%</td>
</tr>
<tr>
<td>Follow</td>
<td>79%</td>
<td>84%</td>
<td>75%</td>
<td>83%</td>
</tr>
</tbody>
</table>

• Overall Accuracy: 74%
Experimental Analysis with HandQA Data

• **Syntactic features**
  – Not helpful
  – Not reliable
    • Typos, grammars, capitalization, punctuation, …

• **Semantic features**
  – More important due to characteristics of data
  – Similar topics in consecutive series
Tree

```
root
  ↓
PATH noun < 0.46
  ↓
PATH verb < 0.31
  ↓
PATH verb > 0.65
  ↓
PATH verb > 0.31
  ↓
PATH verb < 0.65
  ↓
PATH verb > 0.65
```
Summary of Results

• Machine learning approach
  – Flexible, different rules for different data sets
    • Pronoun for TREC; PATH for HandQA
  – Better results
  – Describe the data better

• Semantic similarity
  – PATH is one of the dominating features
Outline

- Introduction
- Previous research work
- A data driven approach
- Results
  - Results on TREC data
  - Results on HandQA data
  - Preliminary contextual information fusion
- Future work
TREC 2004 Data

<target id="2" text="Fred Durst">
  <q id="2.1" type="factoid">
    What is the name of Durst's group?  </q>
  <q id="2.2" type="factoid">
    What record company is he with?  </q>
  <q id="2.3" type="list">
    What are titles of the group's releases?  </q>
  <q id="2.4" type="factoid">
    Where was Durst born?  </q>
</target>
Approaches of TREC Participants

• Search in topic words docs

• Topic words attachment
  • Attach topic words to each question

• Anaphoric replacement
  • Replace pronouns with topic words

• Deep anaphora analysis
  • Trying to find the true referent for pronouns
Is Context Information Useful?

• **Context info doesn’t help.** [Winikoff and Kosseim, 2004]
  - First run: original questions (contextual questions)
  - Second run: manually pronoun replacement (independent questions)
  - Results: not improved
  - Explanation: poor performance of the QA system?
Context Info in DR

• Is context necessary?
  - Let’s use document retrieval

<table>
<thead>
<tr>
<th>The top n documents (n=?)</th>
<th>50</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>Topic words</td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>Topic words+ Question</td>
<td>87%</td>
<td>96%</td>
</tr>
</tbody>
</table>
Use questions in previous turns

<table>
<thead>
<tr>
<th>The top n documents (n=?)</th>
<th>50</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>Topic words <em>(not available in reality)</em></td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>Noun (first question)</td>
<td>81%</td>
<td>92%</td>
</tr>
<tr>
<td>First question</td>
<td>76%</td>
<td>93%</td>
</tr>
<tr>
<td>Topic words + Question</td>
<td>87%</td>
<td>96%</td>
</tr>
<tr>
<td>PN (first question) + Question</td>
<td>77%</td>
<td>92%</td>
</tr>
<tr>
<td>Noun (first question) + Question</td>
<td>84%</td>
<td>94%</td>
</tr>
<tr>
<td>First question + Question</td>
<td>82%</td>
<td>94%</td>
</tr>
</tbody>
</table>
Use questions in previous turns (Cont.)

<table>
<thead>
<tr>
<th>The top n documents (n=?)</th>
<th>50</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questions Only</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>Topic words</td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>Noun</td>
<td>81%</td>
<td>92%</td>
</tr>
<tr>
<td>First question</td>
<td>76%</td>
<td>93%</td>
</tr>
<tr>
<td>Topic words + Question</td>
<td>87%</td>
<td>96%</td>
</tr>
<tr>
<td>PN (first question) + Question</td>
<td>77%</td>
<td>92%</td>
</tr>
<tr>
<td>Noun (first question) + Question</td>
<td>84%</td>
<td>94%</td>
</tr>
<tr>
<td><strong>Incremental Noun</strong></td>
<td><strong>87%</strong></td>
<td><strong>94%</strong></td>
</tr>
<tr>
<td>First question + Question</td>
<td>82%</td>
<td>94%</td>
</tr>
</tbody>
</table>

Incremental Nouns: Nouns in previous questions with the Semantic Similarity PATH > 0.08
Make use of answers in the context

<table>
<thead>
<tr>
<th>Mode</th>
<th>50</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>20%</td>
<td>39%</td>
</tr>
<tr>
<td>Topic words</td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>Noun</td>
<td>81%</td>
<td>92%</td>
</tr>
<tr>
<td>First question</td>
<td>76%</td>
<td>93%</td>
</tr>
<tr>
<td>Topic words + Question</td>
<td>87%</td>
<td>96%</td>
</tr>
<tr>
<td>PN (first question) + Question</td>
<td>77%</td>
<td>92%</td>
</tr>
<tr>
<td>Noun (first question) + Question</td>
<td>84%</td>
<td>94%</td>
</tr>
<tr>
<td>Noun (first question) + Question + Answer</td>
<td>86%</td>
<td>95%</td>
</tr>
<tr>
<td>First question + Question</td>
<td>82%</td>
<td>94%</td>
</tr>
<tr>
<td>First question + Question + Answer</td>
<td>86%</td>
<td>95%</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• Previous research work
• A data driven approach
• Results
  – Results on TREC data
  – Results on HandQA data
• Preliminary contextual information fusion
  ➢ Future work
Future Work

• More research to improve performance
  – Integrate context in QA
  – Evaluate context in QA

• Dialogue-based QA
  – [Small et al. 2004]

• Implement into a QA system