Active Learning for Building a Corpus of Questions for Parsing

Jordi Atserias, Yahoo! Research, Barcelona
Giuseppe Attardi, Università di Pisa
Maria Simi, Università di Pisa
Hugo Zaragoza, Yahoo! Research, Barcelona
Summary

• Introduction and Goals
• Construction of a question corpus
• Experiments
  ▪ Parsing questions / non questions
  ▪ Smartest ways of building the corpus
    • Different criteria, batch size
    • “exploring” active learning
• Conclusions and Further
Motivations

• Accuracy in parsing questions is important
  ▪ question answering, FAQ retrieval, dialogue systems ...  

• Parsers have poor accuracy on questions

• No suitable question specific training resources are available
Need for an specific corpus

- CoNLL 2007
  - only 0.75% are questions, not very representative
  - Annotations are sometimes inconsistent
- Questions have a specific structure
Specific Motivation: Yahoo! Answers

- Several millions of questions collected from users, in several languages
- Yahoo! Answers Collection (Webscope)
  - 4,483,032 questions (and answers)
- Motivation: building a service for question retrieval (Yahoo! Quest available at http://quest.sandbox.yahoo.net)
## English Question corpus

- 800 yahoo! answers questions [relatively clean]
- 500 questions from TREC QA
- PosTagged, revised andParsed with DeSR, revised

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of sentences</th>
<th>Average sentence length</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo! Answers Corpus</td>
<td>800</td>
<td>11.35</td>
<td>9,080</td>
</tr>
<tr>
<td>TREC QA Corpus</td>
<td>500</td>
<td>7.5</td>
<td>3,750</td>
</tr>
<tr>
<td>Question Corpus</td>
<td>1300</td>
<td>9.50</td>
<td>12,830</td>
</tr>
</tbody>
</table>
Research questions

- Q1: how big a corpus of questions should be in order to achieve adequate accuracy?
- Q2: Is a single corpus adequate to analyze both questions and non-questions?
- Q3: Can we minimize the cost of annotating the corpus?

Active learning

- supervised machine learning technique in which the learner is allowed to select the data

Size of data samples

- The smaller the set, the less efficient the process
- Adding training data all at once, no benefit from AL
Experiment Set up

- **Question Corpus (12,830 tokens)**
  - Divided into a base train and base test corpus
- **Base corpus (250,805 tokens)**
  - A sample of CoNLL 2007, without questions
  - Divided into a base train corpus and base test corpus
- **Baseline**
  - Train on a corpus composed of the base train corpus plus random samples of questions of increasing size (0, 100, 200, 300 ... 1000) extracted from the question train corpus
  - For each training corpus:
    - evaluate on the question test (LAS score)
    - evaluate on the base test (LAS score)
  - Repeat with different seeds (5 times), take the average LAS
Q1 (size) & Q2 (helps and no harm)
Random selection

<table>
<thead>
<tr>
<th></th>
<th>base</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>quest LAS</td>
<td>77.20%</td>
<td>81.99%</td>
<td>83.54%</td>
<td>84.59%</td>
<td>85.22%</td>
<td>85.10%</td>
<td>85.23%</td>
<td>85.92%</td>
<td>85.77%</td>
<td>85.81%</td>
<td>86.01%</td>
</tr>
<tr>
<td>base LAS</td>
<td>84.69%</td>
<td>85.73%</td>
<td>84.88%</td>
<td>85.26%</td>
<td>85.34%</td>
<td>85.56%</td>
<td>85.43%</td>
<td>85.32%</td>
<td>85.15%</td>
<td>85.49%</td>
<td>85.63%</td>
</tr>
</tbody>
</table>

Graph showing percentage improvements over various iterations.
Q3: Can minimize annotation effort? Exploring Active learning

- Active learning is an iterative process
- At each step:
  - A learner is trained using the previous model
  - Using a “selection criterion” chooses “interesting” examples from a non-annotated collection (reparse the unannotated data)
  - Manually annotated and added to the training corpus
- If the selection criterion is effective, a much smaller number of examples is needed
Q3: Can minimize annotation effort? Testing selection criteria

- Selection criteria based on the output of the DeSR transition based parser
- Likelihood of sentence parse tree can be computed as the product of the probabilities of all parsing steps
  - LLK: Lowest likelihood of sentence parse tree
  - HLK: Highest likelihood of sentence parse tree
  - LAP: Lowest average probability
  - LNL: Lowest normalized likelihood (likelihood/log(#tokens))
- The sentences in the question training corpus were parsed and then ordered a priori with these criteria.
- Samples of increasing size were tested (as before)
## Evaluation of selection criteria

<table>
<thead>
<tr>
<th></th>
<th>base</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RAND</strong></td>
<td><strong>77.20%</strong></td>
<td>81.99%</td>
<td>83.54%</td>
<td>84.59%</td>
<td>85.22%</td>
<td>85.10%</td>
<td>85.23%</td>
<td>85.92%</td>
<td>85.77%</td>
<td>85.81%</td>
<td>86.01%</td>
</tr>
<tr>
<td><strong>LLK</strong></td>
<td><strong>77.20%</strong></td>
<td>82.87%</td>
<td><strong>85.39%</strong></td>
<td>85.19%</td>
<td>84.99%</td>
<td>85.58%</td>
<td>84.80%</td>
<td>85.58%</td>
<td>86.18%</td>
<td><strong>87.12%</strong></td>
<td>85.74%</td>
</tr>
<tr>
<td><strong>HLK</strong></td>
<td><strong>77.20%</strong></td>
<td>76.84%</td>
<td>77.79%</td>
<td>78.69%</td>
<td>80.19%</td>
<td>82.99%</td>
<td>85.66%</td>
<td>84.29%</td>
<td>84.84%</td>
<td>84.48%</td>
<td>86.14%</td>
</tr>
<tr>
<td><strong>LAP</strong></td>
<td><strong>77.20%</strong></td>
<td>82.71%</td>
<td>83.85%</td>
<td>84.80%</td>
<td>84.60%</td>
<td>86.10%</td>
<td>86.29%</td>
<td>86.33%</td>
<td>85.78%</td>
<td>86.10%</td>
<td>85.70%</td>
</tr>
<tr>
<td><strong>LNL</strong></td>
<td><strong>77.20%</strong></td>
<td>82.20%</td>
<td><strong>85.47%</strong></td>
<td>85.35%</td>
<td>84.17%</td>
<td>85.66%</td>
<td>86.14%</td>
<td>85.19%</td>
<td>85.66%</td>
<td>85.98%</td>
<td><strong>86.92%</strong></td>
</tr>
</tbody>
</table>
Smaller steps

![Graph showing the progression of steps with two lines, one for batch 20 and one for batch 100. The graph shows an increase in percentage from 72% to 88% as steps progress.]
The corpus we have built can be useful for improving the accuracy of parsers in analysing questions.

With a relatively small corpus (about 1000 questions) quite good accuracy can be obtained in parsing questions without hurting the performance on non-question sentences.

By using active learning we can further reduce the cost of building a question corpus.
Future Work

- Building a larger corpus
- Try this approach on other languages
- Explore ML techniques that use unannotated data
Any Question?

Questions and feedback are highly welcome

Thanks for your attention