Good Question! Statistical Ranking for Question Generation

The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics

Michael Heilman and Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
fmheilman,nasmithg@cs.cmu.edu

Reporter: Yi-Ting Huang
Date: 2012/03/09
Outline

- Introduction
- Related Work
- Definitions
- Method:
  - Rule-based question generation
  - Ranking
- Corpora
- Evaluation
- Conclusion and Comments
Introduction

• Our goal is to generate fact-based questions about the content of a given article.

• The top-ranked questions could be filtered and revised by educators, or given directly to students for practice.

Darwin studied how *species* evolve.

Who studied how *species* evolve?

*What did Darwin study how evolve?*
Complex Input Sentences

Lincoln, who was born in Kentucky, moved to Illinois in 1831.

Intermediate Form: Lincoln was born in Kentucky.

Where was Lincoln born?

Step 1: Extraction of Simple Factual Statements

Step 2: Transformation into Questions

Rule-based
Vague and Awkward Questions, etc.

Lincoln, who was born in Kentucky...
Where was Lincoln born?

Lincoln, who faced many challenges...
What did Lincoln face?

Weak predictors:
# proper nouns, WH word, transformations, etc.

Step 1: Extraction of Simple Factual Statements
Step 2: Transformation into Questions
Step 3: Statistical Ranking

Rule-based
Learned from labeled output from steps 1&2
Related work – Question Answering

• While much of the question answering research has focused on retrieval or extraction, models of the transformation from answers to questions have also been developed (Echihabi and Marcu, 2003) with the goal of finding correct answers given a question (e.g., in a source-channel framework).

• In such work on question answering, question generation models are typically not evaluated for their intrinsic quality, but rather with respect to their utility as an intermediate step in the question answering process.
Related work – compare to the other NLG

• Question Generation (QG) is also different from some other tasks related to generation:
  • unlike machine translation (e.g., Brown et al., 1990), the input and output for QG are in the same language, and their length ratio is often far from one to one;
  • unlike sentence compression (e.g., Knight and Marcu, 2000), QG may involve substantial changes to words and their ordering, beyond simple removal of words.
Related work – Question-aid generation

• Some previous research has directly approached the topic of generating questions for educational purposes, but to our knowledge, none has involved statistical models for choosing among output candidates.

• Mitkov et al. (2006) demonstrated that automatic generation and manual correction of questions can be more time-efficient than manual authoring alone.

• Existing QG systems model their transformations from source text to questions with many complex rules for specific question types (e.g., a rule for creating a question *Who did the *Subject* Verb?* from a sentence with *SVO* word order and an object referring to a person), rather than with sets of general rules.
Thus, we employ an overgenerate-and-rank approach, which has been applied successfully in areas such as generation.

We take a rule-based approach in order to leverage this linguistic knowledge.

- The characteristics of question generation are (arguably) difficult to learn from corpora, but they have been studied extensively in linguistics.

Since large datasets of the appropriate domain, style, and form of questions are not available to train our ranking model, we learn to rank from a relatively small, tailored dataset of human-labeled output from our rule-based system.
Definitions

source sentence  
Darwin studied how species evolve.

question phrase  
Who studied how species evolve? 🤚

answer phrase  
Darwin
Implementation

- We use phrase structure parses from Stanford Parser.
- We encode transformations in the Tregex tree searching language, and Tsurgeon, a tree manipulation language.
- We use BBN Indentifinder to find entity labels, and map these to WH words.
  - PERSON -> Who
  - LOCATION -> Where
  - etc.
Example Tregex Rule

Constraint: Phrases dominated by a clause with a WH-complementizer cannot undergo movement.

\[
\text{SBAR} < /\text{^WH}.*P$/> \ll \text{NP} | \text{ADJP} | \text{VP} | \text{ADVP} | \text{PP}=\text{unmv}
\]

```
Darwin studied how species evolve.
```

"<" denotes dominance

Step 1 Sentence Simplification

- An analysis of the question transducer’s output when applied directly to sentences from source texts revealed that complex sentences often lead to unnatural or even senseless questions.
- Various simplifying transformations are performed in step 1 to remove phrase types.
The set of rules

- To remove phrase types:

<table>
<thead>
<tr>
<th>Description of Transformation</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence-initial conjunctions are removed by deleting conj.</td>
<td>ROOT &lt; (S &lt; CC=conj)</td>
</tr>
<tr>
<td>Sentence-initial adjunct phrases are removed by deleting adjunct. (A nearly identical rule</td>
<td>ROOT &lt; (S &lt; (/[^,]/=adjunct $.. (//, $.. VP)))</td>
</tr>
<tr>
<td>deletes commas following these adjuncts.)</td>
<td></td>
</tr>
<tr>
<td>Appositives are removed by deleting app, lead, and trail. (A nearly identical rule deletes</td>
<td>SBAR</td>
</tr>
<tr>
<td>parenthetical phrases.)</td>
<td></td>
</tr>
</tbody>
</table>

- Questions can be produced about much of this embedded content if we extract declarative sentences from finite clauses, relative clauses, appositives, participial phrases. The other questions will not be generated by the question transducer.
During the Gold Rush years in northern California, Los Angeles became known as the "Queen of the Cow Counties" for its role in supplying beef and other foodstuffs to hungry miners in the north.
Step 2 Question Transformation

Step 1: Extraction of Simple Factual Statements

Step 2: Transformation into Questions

Step 3: Statistical Ranking

Rule-based

Learned from labeled output from steps 1&2

Declarative Sentence

Mark Unmovable Phrases

Choose Answer Phrase *

Generate Possible Question Phrase *

(Decompose Main Verb)

(Invert Subject and Auxiliary)

Insert Question Phrase

Perform Post-processing

Question
Mark unremovable phrases

- I bought the book that inspired Bob.
- Who did I buy the book that inspired? 😞
- To whom did I give the book? 😞
- Whom did I give the book to? 😞

Choose answer phrases

<table>
<thead>
<tr>
<th>Question Phrase</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who</td>
<td>a noun phrase whose head is labeled PERSON, PER_DESC, or ORGANIZATION</td>
</tr>
<tr>
<td>Where</td>
<td>a noun phrase whose head is labeled LOCATION, or is a prepositional phrase with certain prepositions (in, at, on, over) whose head is labeled LOCATION</td>
</tr>
<tr>
<td>When</td>
<td>a noun phrase whose head is labeled DATE or TIME, or is a prepositional phrase with certain prepositions (in, at, on, over) whose head is labeled DATE or TIME</td>
</tr>
<tr>
<td>How much</td>
<td>a noun phrase with a quantifier phrase or word (“QP” or “CD”) and whose head is labeled MONEY</td>
</tr>
<tr>
<td>A preposition followed by any of the above</td>
<td>a prepositional phrase whose object is a noun phrase that satisfies one of the above conditions</td>
</tr>
</tbody>
</table>
• John saw Mary. → Who did John see?  
   → Saw John Mary?  
   → John did see Mary. (Decomposition of the main verb) 
   → did John see Mary. (Insert Subject-Auxiliary) 
   → Who did John see Mary. (Insert Question Phrase) 
   → Who did John see Mary? (Post-processing)
Los Angeles became known as the "Queen of the Cow Counties" for its role in supplying beef and other foodstuffs to hungry miners in the north.
Step 3 Ranking

What did Los Angeles become known as the "Queen of the Cow Counties" for?

Question Ranking

1. What became known as...?
2. What did Los Angeles become known as the "Queen of the Cow Counties" for?
3. Whose role in supplying beef...?
4. ...

Step 1: Extraction of Simple Factual Statements
Step 2: Transformation into Questions
Step 3: Statistical Ranking

Rule-based

Learned from labeled output from steps 1&2
The ranking model

- logistic regression model (Collins, 2000)

### Table: Question Deficiencies and Description

<table>
<thead>
<tr>
<th>Question Deficiency</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ungrammatical</td>
<td>The question is not a valid English sentence. (e.g., <em>In what were nests excavated exposed to the sun?</em> from ... <em>eggs are usually laid ...</em> in nests excavated in pockets of earth exposed to the sun. This error results from the incorrect attachment by the parser of <em>exposed to the sun</em> to the verb phrase headed by <em>excavated)</em></td>
<td>14.0</td>
</tr>
<tr>
<td>Does not make sense</td>
<td>The question is grammatical but indecipherable. (e.g., <em>Who was the investment?</em>)</td>
<td>20.6</td>
</tr>
<tr>
<td>Vague</td>
<td>The question is too vague to know exactly what it is asking about, even after reading the article (e.g., <em>What do modern cities also have?</em> from ... <em>but modern cities also have many problems</em>).</td>
<td>19.6</td>
</tr>
<tr>
<td>Obvious answer</td>
<td>The correct answer would be obvious even to someone who has not read the article (e.g., a question where the answer is obviously the subject of the article).</td>
<td>0.9</td>
</tr>
<tr>
<td>Missing answer</td>
<td>The answer to the question is not in the article.</td>
<td>1.4</td>
</tr>
<tr>
<td>Wrong WH word</td>
<td>The question would be acceptable if the WH phrase were different (e.g., a <em>what</em> question with a person’s name as the answer).</td>
<td>4.9</td>
</tr>
<tr>
<td>Formatting</td>
<td>There are minor formatting errors (e.g., with respect to capitalization, punctuation).</td>
<td>8.9</td>
</tr>
<tr>
<td>Other</td>
<td>The question was unacceptable for other reasons.</td>
<td>1.2</td>
</tr>
<tr>
<td>None</td>
<td>The question exhibits none of the above deficiencies and is thus acceptable.</td>
<td>27.3</td>
</tr>
</tbody>
</table>

Features

- Length Features (3i+24b)
- WH Words (9b)
- Negation (1b)
- N-Gram Language Model Features (6r)
- Grammatical Features (23i+75b)
  - the numbers of POS tag, NPs, VPs, etc.
- Transformations (8b)
  - e.g., removal of appositives
- Vagueness (3i+15b)
  - Counts of NPs headed by common nouns and with no modifiers
    - 1 for “the president”
    - 0 for “Abraham Lincoln” or “the U.S. president during the Civil War”
## Corpora

<table>
<thead>
<tr>
<th></th>
<th>English Wikipedia</th>
<th>Simple English Wikipedia</th>
<th>Wall Street Journal (PTB Sec. 23)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texts</td>
<td>14</td>
<td>18</td>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>Questions</td>
<td>1,448</td>
<td>1,313</td>
<td>474</td>
<td>3,235</td>
</tr>
</tbody>
</table>

- Training
  - 2,807 questions
  - 36 texts
- Testing
  - 428 questions
  - 6 texts

- Params. are estimated by optimizing $L_2$ regularized conditional log-likelihood.
- We use a variant of Newton’s method.
Rating

- Fifteen native English-speaking university students rated output from the overgeneration steps 1&2.
  - No deficiencies: 
  - Any deficiencies: 
- An inter-rater agreement of $K = 0.42$

<table>
<thead>
<tr>
<th>Question Deficiency</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ungrammatical</td>
<td>The question is not a valid English sentence. (e.g., <em>In what were nests excavated exposed to the sun?</em> from ... <em>eggs are usually laid ... in nests excavated in pockets of earth exposed to the sun</em>. This error results from the incorrect attachment by the parser of <em>exposed to the sun</em> to the verb phrase headed by <em>excavated</em>)</td>
<td>14.0</td>
</tr>
<tr>
<td>Does not make sense</td>
<td>The question is grammatical but indecipherable. (e.g., <em>Who was the investment?</em>)</td>
<td>20.6</td>
</tr>
<tr>
<td>Vague</td>
<td>The question is too vague to know exactly what it is asking about, even after reading the article (e.g., <em>What do modern cities also have?</em> from ... <em>but modern cities also have many problems</em>).</td>
<td>19.6</td>
</tr>
<tr>
<td>Obvious answer</td>
<td>The correct answer would be obvious even to someone who has not read the article (e.g., a question where the answer is obviously the subject of the article).</td>
<td>0.9</td>
</tr>
<tr>
<td>Missing answer</td>
<td>The answer to the question is not in the article.</td>
<td>1.4</td>
</tr>
<tr>
<td>Wrong WH word</td>
<td>The question would be acceptable if the WH phrase were different (e.g., <em>a what question with a person’s name as the answer</em>).</td>
<td>4.9</td>
</tr>
<tr>
<td>Formatting</td>
<td>There are minor formatting errors (e.g., with respect to capitalization, punctuation).</td>
<td>8.9</td>
</tr>
<tr>
<td>Other</td>
<td>The question was unacceptable for other reasons.</td>
<td>1.2</td>
</tr>
<tr>
<td>None</td>
<td>The question exhibits none of the above deficiencies and is thus acceptable.</td>
<td>27.3</td>
</tr>
</tbody>
</table>
Evaluation

- Metric:
  the percentage of test set questions labeled as acceptable.

- Baseline:
  - Ranker with all features
  - Ranker with surface features
  - Ranker with features only about question
  - Random
  - Oracle: if all questions that were labeled acceptable were ranked higher than all questions that were labeled unacceptable
Noisy at top ranks.

Ranking Results

- Oracle
- All Features
- Surface Features
- Expected Random

Pct. Rated Acceptable vs Number of Top-Ranked Questions

All Features performed significantly better than Surface Features ($p < .05$).
## Ablation Experiments

<table>
<thead>
<tr>
<th>Features</th>
<th>#</th>
<th>Top 20%</th>
<th>Top 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>187</td>
<td>52.3</td>
<td>40.8</td>
</tr>
<tr>
<td>All – Length</td>
<td>160</td>
<td>52.3</td>
<td>42.1</td>
</tr>
<tr>
<td>All – WH</td>
<td>178</td>
<td>50.6</td>
<td>39.8</td>
</tr>
<tr>
<td>All – Negation</td>
<td>186</td>
<td>51.7</td>
<td>39.3</td>
</tr>
<tr>
<td>All – Lang. Model</td>
<td>181</td>
<td>51.2</td>
<td>39.9</td>
</tr>
<tr>
<td>All – Grammatical</td>
<td>69</td>
<td>43.2</td>
<td>38.7</td>
</tr>
<tr>
<td>All – Transforms</td>
<td>179</td>
<td>46.5</td>
<td>39.0</td>
</tr>
<tr>
<td>All – Vagueness</td>
<td>169</td>
<td>48.3</td>
<td>41.5</td>
</tr>
<tr>
<td>All – Histograms</td>
<td>53</td>
<td>49.4</td>
<td>39.8</td>
</tr>
<tr>
<td>Surface</td>
<td>43</td>
<td>39.5</td>
<td>37.6</td>
</tr>
<tr>
<td>Question Only</td>
<td>91</td>
<td>41.9</td>
<td>39.5</td>
</tr>
<tr>
<td>Random</td>
<td>-</td>
<td>27.3</td>
<td>27.3</td>
</tr>
<tr>
<td>Oracle</td>
<td>-</td>
<td>100.0</td>
<td>87.3</td>
</tr>
</tbody>
</table>
Conclusion

- Overgeneration and ranking for QG.
  - Rules encode linguistic knowledge
  - Statistical ranker captures trends not easily encoded with rules
- Statistical ranking improved top-ranked output.
Comments

• Contribution:
  • Rules encode linguistic knowledge
  • Statistical ranker captures trends not easily encoded with rules
• Idea 1:
  • Reading comprehension items selection
• Idea 2:
  • Distroactors ranking in vocabulary:
  • Questions ranking in vocabulary:
    • word difficulty based on word lists
    • mistakes an examinee made
    • word estimated difficulty (tf/idf)
Q&A

Generated from our paper’s abstract:

- Which challenge do we address?
- Who use manually written rules to perform a sequence of general purpose syntactic transformations to turn declarative sentences into questions?
- Is our approach to overgenerate questions, then rank them?
- What kind of regression model are these questions then ranked by?
- What do experimental results show that ranking nearly doubles?
- What kind of results show that ranking nearly doubles the percentage of questions rated as acceptable by annotators ranked 20% of questions?