Good Question!
Statistical Ranking for Question Generation

Michael Heilman and Noah A. Smith
The Goal

- Input: educational text
- Output: quiz
The Goal

- Input: educational text
- Output: quiz
- Output: ranked list of candidate questions to present to a teacher

Text-to-text generation

Knight & Marcu, 00; Clarke, 06 (Compression); Barzilay & McKeown, 05 (Sentence Fusion); Callison-Burch, 07 (Paraphrase Generation); *inter alia*
Our Approach

- Sentence-level factual questions
- Acceptable (e.g., grammatical) questions
- QG as a series of sentence structure transformations
Outline

- Challenges in Question Generation (QG)
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
Constraints on WH movement

*Darwin studied how *species* evolve.*

*Who studied how species evolve?*

*What did Darwin study how evolve?*

- WH movement is well studied.
- We encode this linguistic knowledge with rules.
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
Connections to Prior Work on QG

Most prior work:

- Sentence-level factual questions
- Syntactic rules for transformation or extraction
- Generation in a single step

Contributions:

- Multi-step framework
- Ranking model learned from labeled output
- QG evaluation methodology with broad-domain corpora

Overgeneration and Ranking for NLG:
Langkilde & Knight 98; Walker et al., 01
Outline

- Challenges in QG
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
Implementation Details

- We use BBN Indentifinder to find entity labels, and map these to WH words.
  - PERSON -> Who
  - LOCATION -> Where
  - etc.

- We use phrase structure parses from Stanford Parser.

- We encode transformations in the Tregex tree searching language.

Bikel et al., 99
Klein & Manning, 03
Levy & Andrew, 06
Example Tregex Rule

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

```
SBAR < /^[WH.]*P$/ << NP|ADJP|VP|ADVP|PP=unmv
```

```
...  
    SBAR
       \  
        WHAVP
           \ 
            WRB
               \  
                S
                   \ 
                    NP
                      \  
                        VP
```

* What did Darwin study how _ evolve?

Darwin studied how species evolve.

Outline

- Challenges in QG
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
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.
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.

Did Los Angeles become known as the "Queen of the Cow Counties" for
(Answer Phrase: its role in…)

Los Angeles did become known as the "Queen of the Cow Counties" for
(Answer Phrase: its role in…)

Los Angeles became known as the "Queen of the Cow Counties" for (Answer Phrase: its role in…)

Subject Auxiliary Inversion

Main Verb Decomposition

Answer Phrase Selection
Did Los Angeles become known as the "Queen of the Cow Counties" for
(Answer Phrase: its role in...)

Movement and Insertion of Question Phrase

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. ...
Outline

- Challenges in QG
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
Rating Questions

- We use rated questions to...
  - Learn a ranking model
  - Evaluate our system
Sources of Data

- Existing datasets of questions?
  - Not focused on sentence-level facts
  - Lack negative examples
  - Noisy (e.g., Yahoo questions)
  - Relatively small

- Tailored data set: annotators rated output from the overgeneration steps 1&2.
Rating Scheme

- 8 possible deficiencies
  - ungrammatical, vague, wrong WH word,…
- Binary rating for each

- No deficiencies: 
- Any deficiencies: 
- “Moderate” agreement ($\kappa = .42$)
## 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
Outline

- Challenges in QG
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
Ranking Model

- Logistic Regression

\[ y \in \{ \text{👍}, \text{👎} \} \]

- Params. are estimated by optimizing L^2 regularized conditional log-likelihood.
- We use a variant of Newton’s method.

To rank, sort by \( P(\text{👍}) \)
Surface Features

- WH words in question
- Negation words in question
- Language model probabilities
- Sentence lengths

Separate features for question, source sentence, answer phrase
Features based on Syntactic Analysis

- **Grammatical categories**
  - Numbers of POS tags, NPs, VPs, etc.

- **Transformations**
  - E.g., extracted from relative clause

- **“Vague NP”**
  - Counts of NPs headed by common nouns and with no modifiers
  - 1.0 for “the president”
  - 0.0 for “Abraham Lincoln” or “the U.S. president during the Civil War”
Outline

- Challenges in QG
- Implementation Details
- Step-by-Step Example
- Rating Questions
- Ranking Model
- Experiments
Evaluation Metric

Percentage of top-ranked test set questions that were rated acceptable (👍)
Rankers & Baselines

- Ranker with all features
- Ranker with surface features
  - only sentence lengths, WH words, negation, language model log probabilities.
- Expected random (i.e., no ranking)
- Oracle
Noisy at top ranks.

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

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>% Acceptable in Top Ranked Fifth</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
<td>52.3</td>
</tr>
<tr>
<td>All – Length</td>
<td>52.3</td>
</tr>
<tr>
<td>All – Negation</td>
<td>51.7</td>
</tr>
<tr>
<td>All – Lang. Model</td>
<td>51.2</td>
</tr>
<tr>
<td>All – WH</td>
<td>50.6</td>
</tr>
<tr>
<td>All – Vagueness</td>
<td>48.3</td>
</tr>
<tr>
<td>All – Transforms</td>
<td>46.5</td>
</tr>
<tr>
<td>All – Grammatical</td>
<td>43.2</td>
</tr>
</tbody>
</table>
Conclusions

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

- Statistical ranking improved top-ranked output.
Questions?

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?