

# Good Question!

## Statistical Ranking for Question Generation

*Michael Heilman and Noah A. Smith*



**Carnegie Mellon**



**pier** @ Carnegie Mellon  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# The Goal

- Input: educational text
- Output: quiz



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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*



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Our Approach

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



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Outline

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



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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.

Ross, 67;  
Chomsky, 77;  
*inter alia*

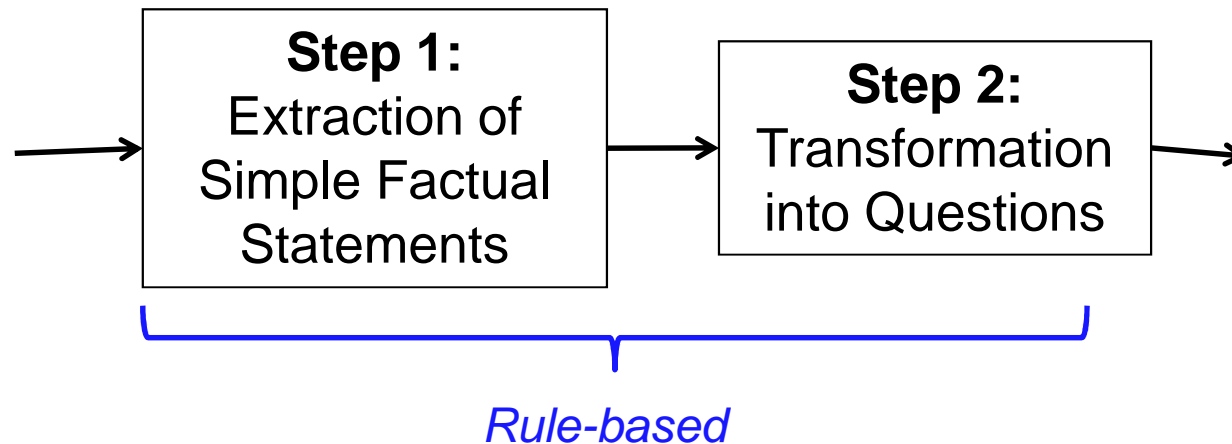


# Complex Input Sentences

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

**Intermediate Form:** *Lincoln was born in Kentucky.*

*Where was Lincoln born?*



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Vague and Awkward Questions, etc.

*Lincoln, who was born in Kentucky...*

*Where was Lincoln born?*



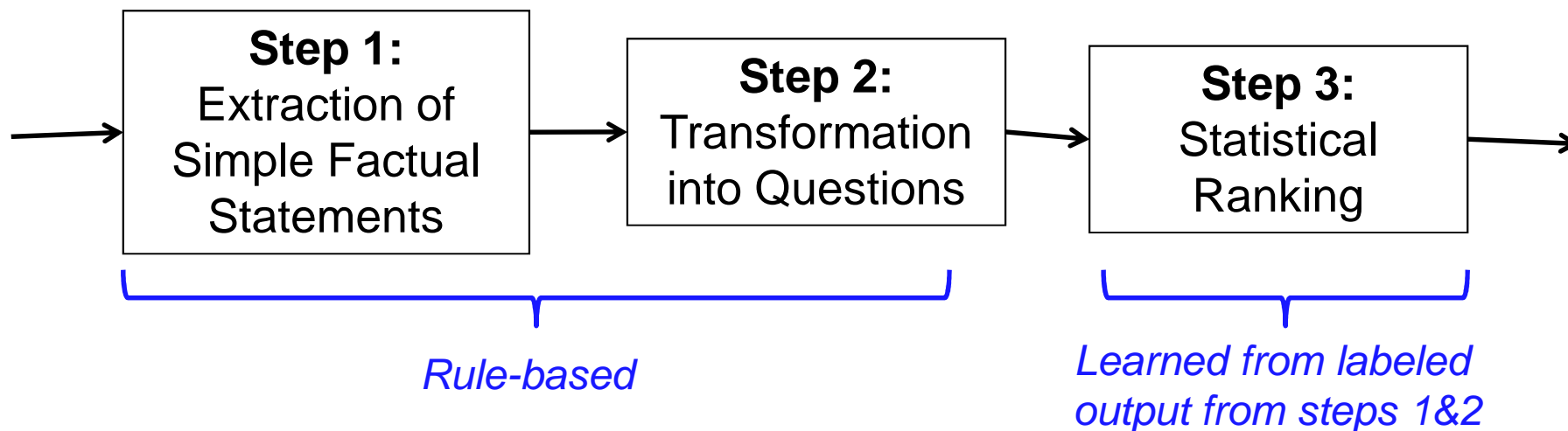
*Lincoln, who faced many challenges...*

*What did Lincoln face?*



**Weak predictors:**

# proper nouns,  
WH word,  
transformations,  
etc.



# Connections to Prior Work on QG

## ■ Most prior work:

Mitkov & Ha, 03; Kunichika *et al.*, 04;  
Gates, 08; *inter alia*

- 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



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Outline

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



CarnegieMellon

**pier**  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Implementation Details

- We use BBN Indentifinder to find entity labels, and map these to WH words.
  - PERSON -> Who Bikel *et al.*, 99
  - LOCATION -> Where
  - etc.
- We use phrase structure parses from Stanford Parser. Klein & Manning, 03
- We encode transformations in the Tregex tree searching language. Levy & Andrew, 06



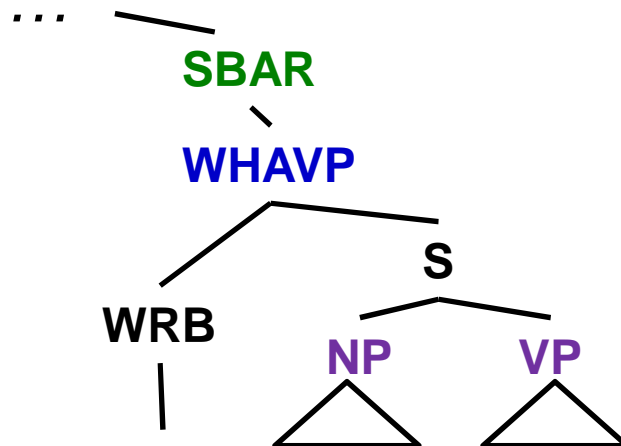
Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Example Tregex Rule

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

**SBAR** < /<sup>^</sup>WH.\*P\$/ << NP | ADJP | VP | ADVP | PP=unmv



“<” denotes dominance

\* *What did Darwin study how \_ evolve?*

*Darwin studied how species evolve.*

**More details on rules in technical report:** M. Heilman and N. A. Smith. 2009. Question Generation via Overgenerating Transformations and Ranking.

# Outline

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

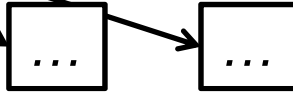


Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH



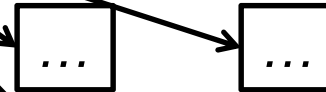
(other candidates)



Preprocessing

*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.*

**Extraction of Simplified Factual Statements**



*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.*

**Answer Phrase Selection**



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

**Main Verb Decomposition**

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

**Subject Auxiliary Inversion**

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

*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



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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

Potential future work



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



CarnegieMellon

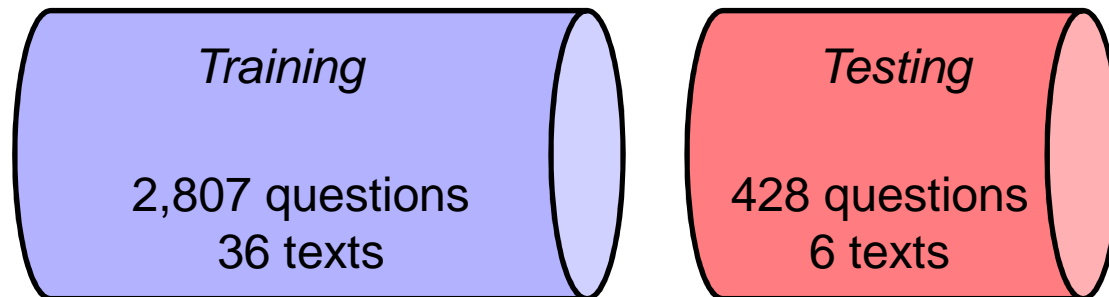
pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Rating Scheme

- 8 possible deficiencies
  - ungrammatical, vague, wrong WH word,...
- Binary rating for each
- No deficiencies: 
- Any deficiencies: 
- “Moderate” agreement ( $\kappa = .42$ )

# Corpora

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



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Outline

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



Carnegie Mellon

**pier**  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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.

le Cessie & Houwelingen, 97

## ■ 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



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# Evaluation Metric



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



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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**



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

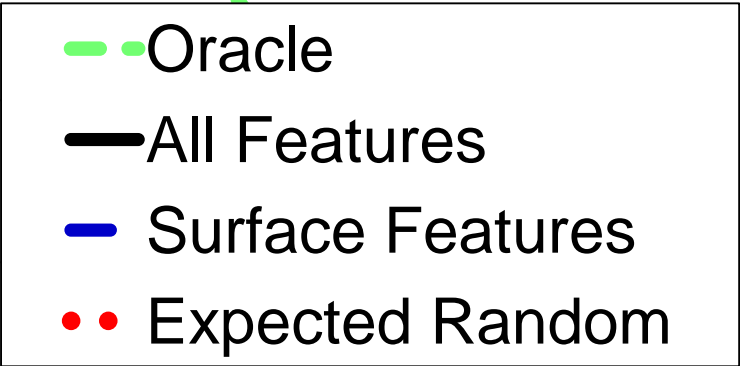
Noisy at top ranks.

# Ranking Results

Testing

Pct. Rated Acceptable

60%  
50%  
40%  
30%  
20%



0 100 200 300 400

Number of Top-Ranked Questions

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

# Ablation Experiments

Feature Set	% Acceptable in Top Ranked Fifth
All Features	52.3
All – Length	52.3
All – Negation	51.7
All – Lang. Model	51.2
All – WH	50.6
All – Vagueness	48.3
All – Transforms	46.5
All – Grammatical	43.2



Carnegie Mellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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.



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH

# 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?



CarnegieMellon

pier  
PROGRAM FOR INTERDISCIPLINARY EDUCATION RESEARCH