

# 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](mailto: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?* 👍

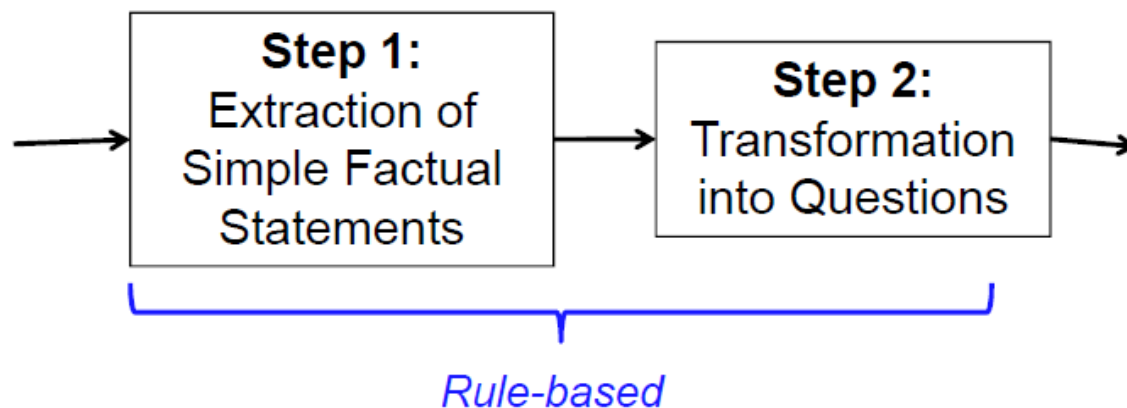
*\*What did Darwin study how evolve?* 👎

# 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?*



# Vague and Awkward Questions, etc.

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

*Where was Lincoln born?*



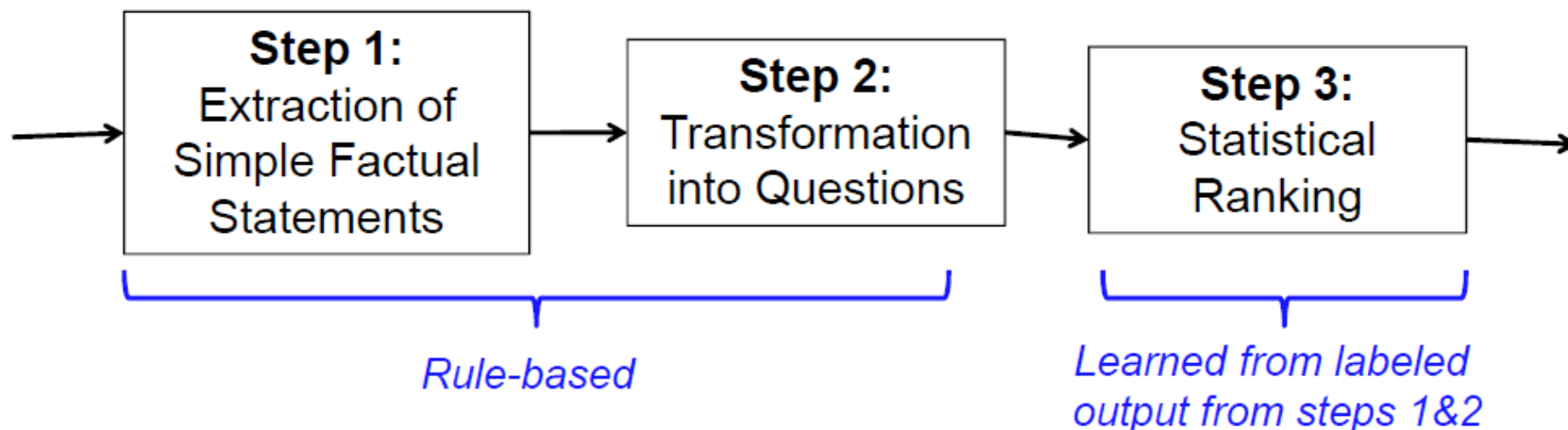
*Lincoln, who faced many challenges...*

*What did Lincoln face?*



**Weak predictors:**

# proper nouns,  
WH word,  
transformations,  
etc.



# 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.



# Introduction (cont.)

- 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*

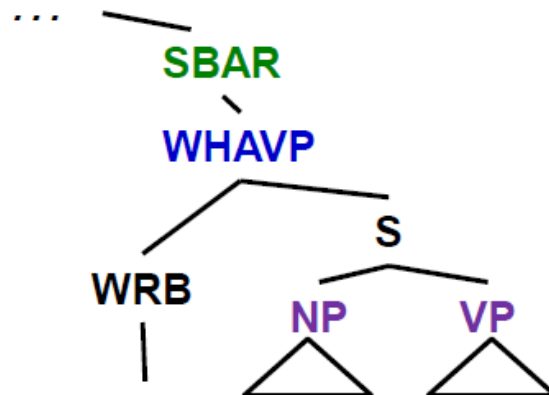
# Implementation

- 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.

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



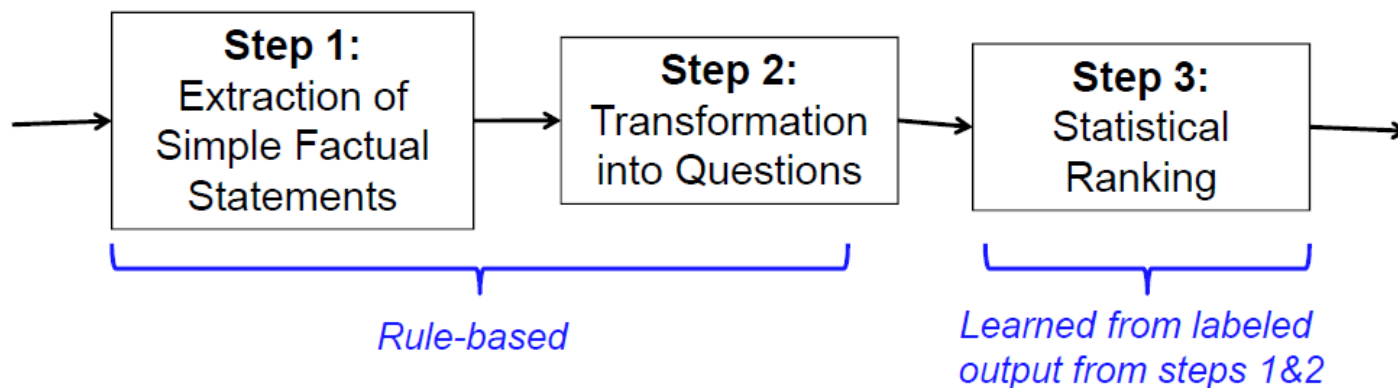
“<” denotes dominance

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

*Darwin studied how species evolve.*

# 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

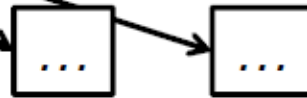
- To remove phrase types:

Description of Transformation	Expression
Sentence-initial conjunctions are removed by deleting <code>conj</code> .	ROOT < (S < CC=conj)
Sentence-initial adjunct phrases are removed by deleting <code>adjunct</code> . (A nearly identical rule deletes commas following these adjuncts.)	ROOT < (S < ([^,]/=adjunct \$.. ([/,/ \$.. VP)))
Appositives are removed by deleting <code>app</code> , <code>lead</code> , and <code>trail</code> . (A nearly identical rule deletes parenthetical phrases.)	SBAR VP NP=app \$, /,/=lead \$. /,/=trail !\$ CC !\$ CONJP

- 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.



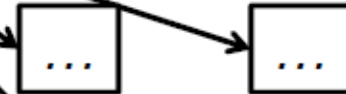
(other candidates)



**Preprocessing**

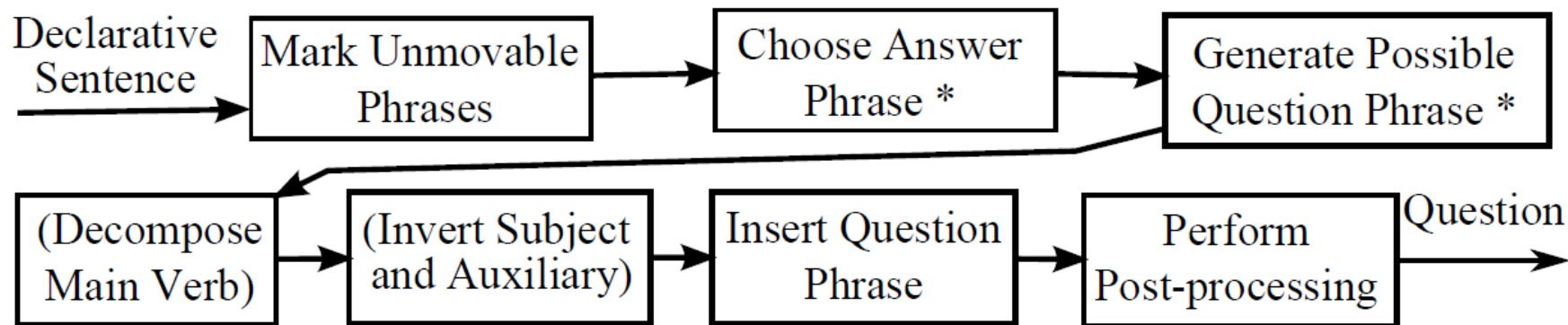
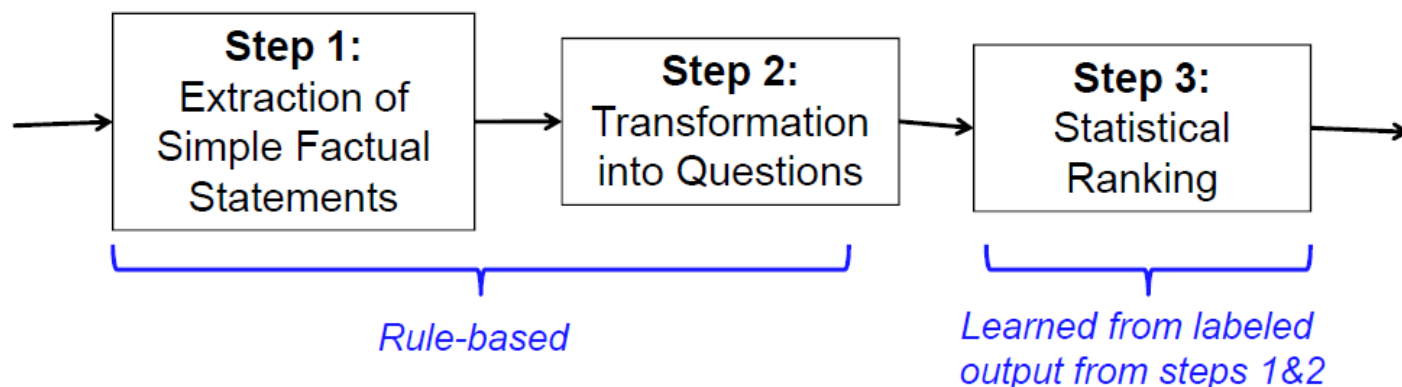
*During the Gold Rush years in northern California, Los Angeles became known as the "Queen of the Cow Counties" for its role in supplying beef and other foodstuffs to hungry miners in the north.*

**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.*

# Step 2 Question Transformation





# 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

Question Phrase	Conditions
Who	a noun phrase whose head is labeled PERSON, PER_DESC, or ORGANIZATION
Where	a noun phrase whose head is labeled LOCATION, or is a prepositional phrase with certain prepositions ( <i>in, at, on, over</i> ) whose head is labeled LOCATION
When	a noun phrase whose head is labeled DATE or TIME, or is a prepositional phrase with certain prepositions ( <i>in, at, on, over</i> ) whose head is labeled DATE or TIME
How much	a noun phrase with a quantifier phrase or word (“QP” or “CD”) and whose head is labeled MONEY
A preposition followed by any of the above	a prepositional phrase whose object is a noun phrase that satisfies one of the above conditions

• *John saw Mary.* → *Who did John see?* 😊

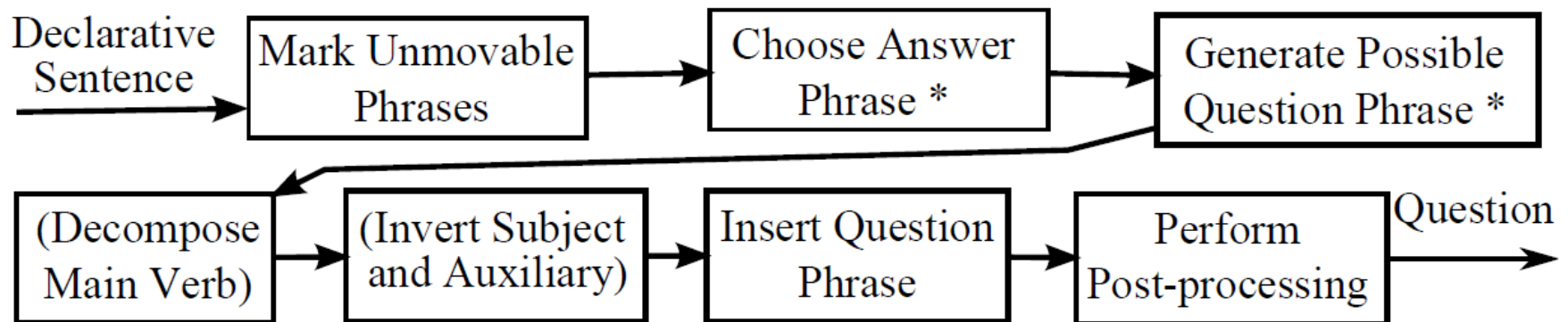
→ *Saw John Mary?* 😞

→ *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.

### 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...)

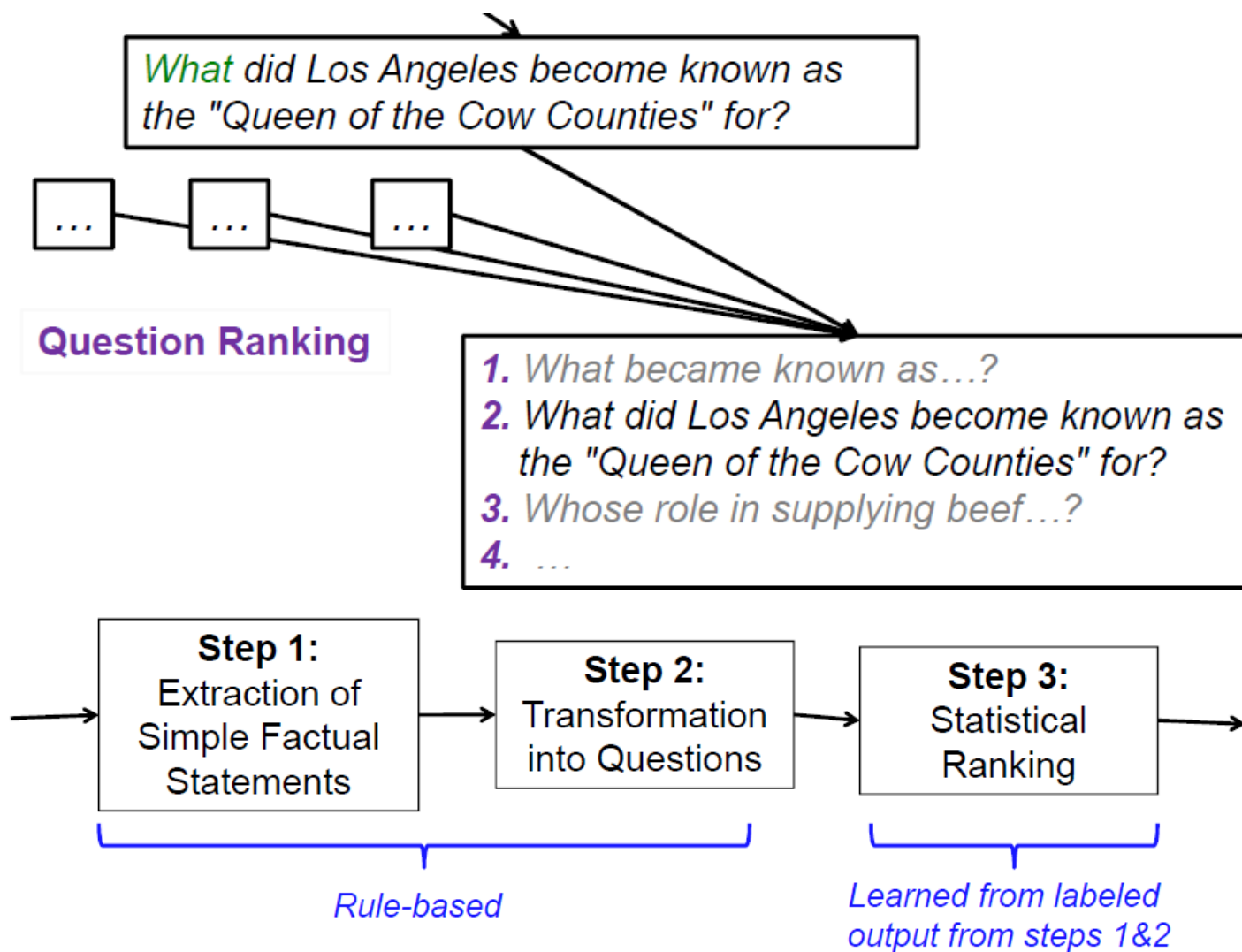
### Subject Auxiliary Inversion

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

### Movement and Insertion of Question Phrase

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

# Step 3 Ranking



# The ranking model

- logistic regression model (Collins, 2000)

➔ { No deficiencies: 👍  
Any deficiencies: 👎 To rank, sorted by  $P(\text{👍})$

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

# Features

{  
i: Integer  
b: boolean  
r: real value

Histogram { 0,1,2,3,4  
Threshold { 0,4,8,12,16,20,24,28

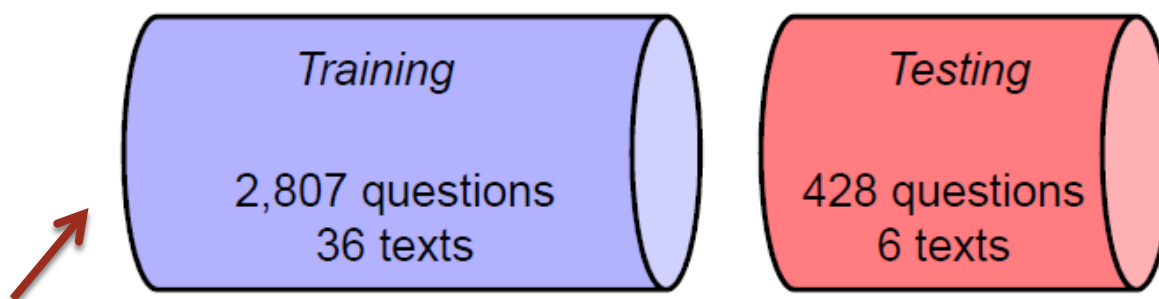
- Length Features (3i+24b)
- WH Words (9b)
- Negation (1b)
- N-Gram Language Model Features (6r)
- Grammatical Features (23i+75b)
  - the numbers of POS tag, NPs, VPs, etc.
- Transformations (8b)
  - e.g., removal of appositives
- Vagueness (3i+15b)
  - Counts of NPs headed by common nouns and with no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

Separate features for question,  
source sentence, answer phrase

Surface feature



# 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



- 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  $K = 0.42$

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



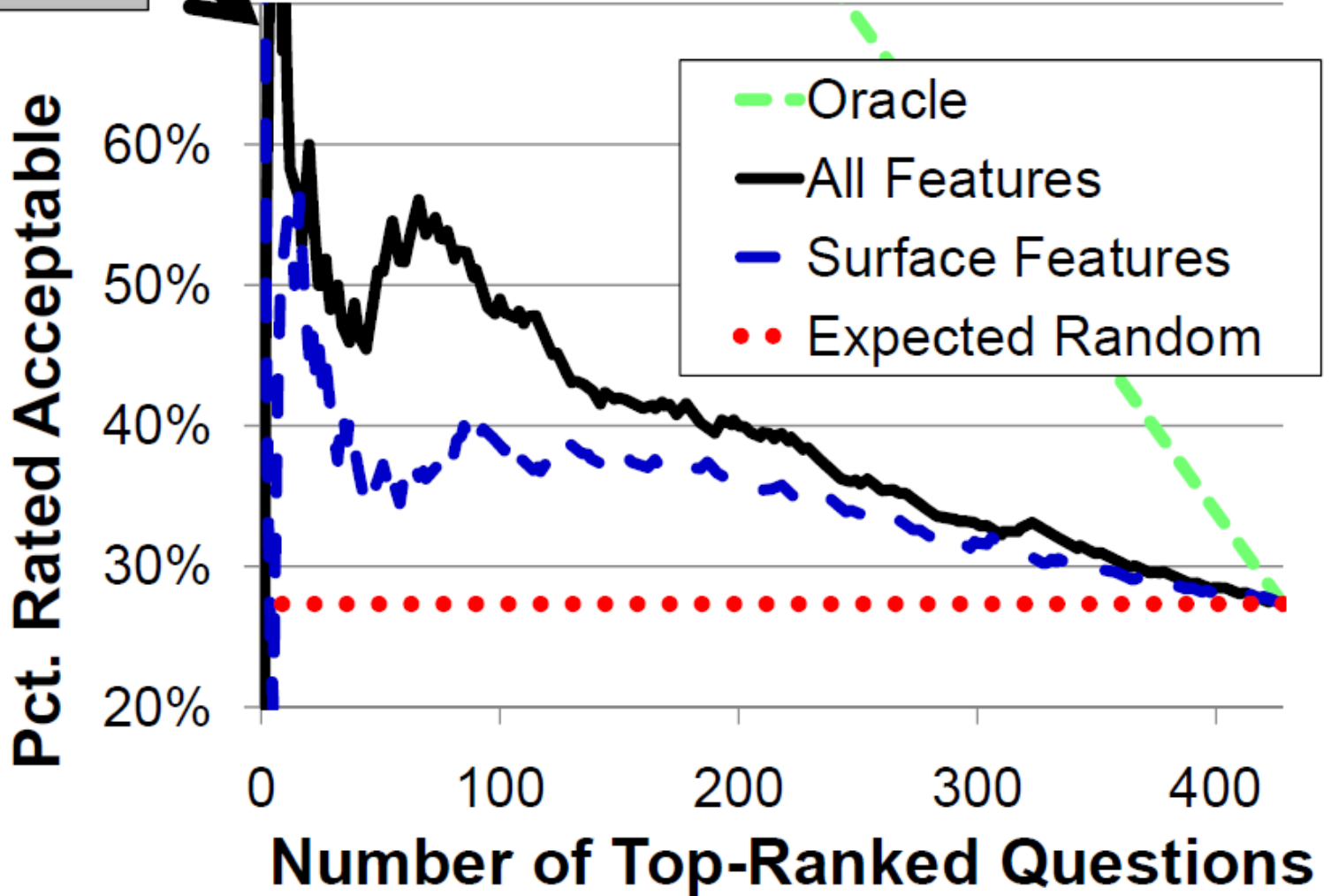
# Evaluation

- 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

Testing



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

# Ablation Experiments

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

# 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:
  - Distractors 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?