

Using Natural Language Processing to Develop Instructional Content

Michael Heilman

Language Technologies Institute

Carnegie Mellon University

REAP Collaborators:

Maxine Eskenazi, Jamie Callan,
Le Zhao, Juan Pino, et al.

Question Generation

Collaborator:
Noah A. Smith



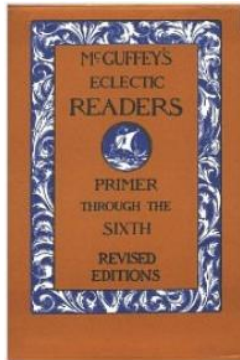
Motivating Example

Situation: Greg, an English as a Second Language (ESL) teacher, wants to find a text that...

- is in grade 4-7 reading level range,
- uses specific target vocabulary words from his class,
- discusses a specific topic, international travel.

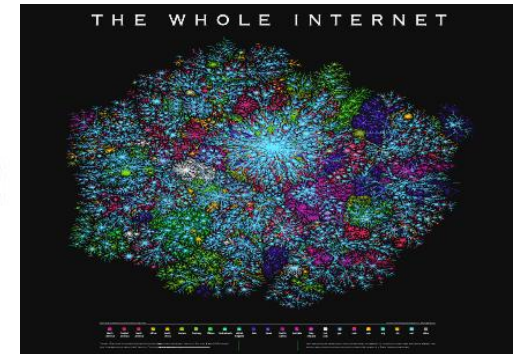
Sources of Reading Materials

Textbook



Internet, etc.

WIKIPEDIA



The New York Times Google books

Google news



Why aren't teachers using Internet text resources *more*?

- Teachers are smart
- Teachers work hard.
- Teachers are computer-savvy.
- Using new texts raises some important challenges...

Why aren't teachers using Internet text resources *more*?

My claim: teachers need better tools...

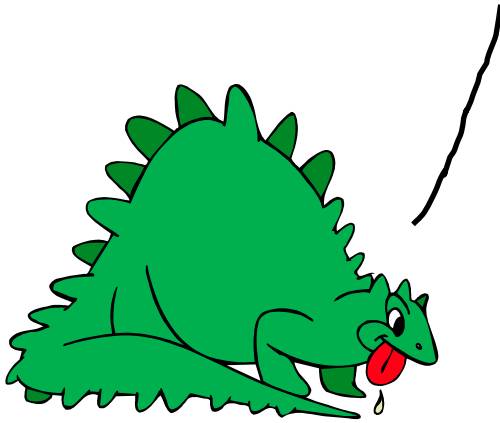
- to find relevant content,
- to create exercises and assessments.

Natural Language Processing (NLP)
can help.

Working Together

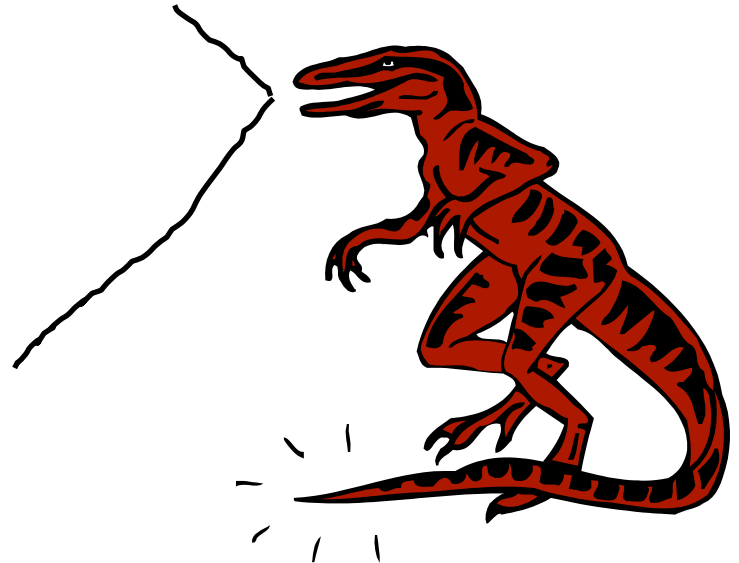
	NLP	Educators	NLP + Educators
Rate of text analysis	Fast	Slow	Fast
Error rate when creating educational content	High	Low	Low

So, what was the talk about?



It was about how tailored applications of Natural Language Processing (NLP) can help educators create instructional content.

It was also about the challenges of using NLP in applications.



Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation (QG)
- Concluding Remarks

Textbooks	New Resources
Fixed, limited amount of content.	Virtually unlimited content on various topics.

Textbooks	New Resources
Fixed, limited amount of content.	Virtually unlimited content on various topics.
Filtered for reading level, vocabulary, etc.	Unfiltered.

Textbooks	New Resources	
Fixed, limited amount of content.	Virtually unlimited content on various topics.	
Filtered for reading level, vocabulary, etc.	Unfiltered.	REAP Search Tool
Include practice exercises and assessments.	No exercises.	Automatic Question Generation

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
 - Motivation
 - NLP components
 - Pilot study
- Question Generation
- Concluding Remarks

REAP Collaborators:

Maxine Eskenazi, Jamie Callan,
Le Zhao, Juan Pino, et al.

The Goal

- To help English as a Second Language (ESL) teachers find reading materials
 - For a particular curriculum
 - For particular students

Back to the Motivating Example

- ***Situation***: Greg, an ESL teacher, wants to find texts that...
 - Are in grade 4-7 reading level range,
 - Use specific target vocabulary words from class,
 - Discuss a specific topic, international travel.
- ***First Approach***: Searching for “international travel” on a commercial search engine...

Typical Web Search Result

vayama™

[Latest Travel Alerts From Vayama](#) *international travel solved*

[flights](#) [hotels](#) [cars](#) [deals](#) [travel extras](#) [my trip](#) [about us](#) [my account](#) [help](#)

find a flight

Round Trip One Way
 Multiple Cities

From: _____

Departing: dd-mon-yyyy Time: Anytime

To: _____

Returning: dd-mon-yyyy Time: Anytime

Adults: 1 Children: 0 Infants: 0 Students: 0
12+ 2-11 0-2

Cabin/Class: Economy / Coach

Carrier Preferences: _____

RESET SEARCH

Win a \$500 Travel Coupon

Join Our Deals Newsletter

Privacy by [SafeSubscribe™](#)

Sign up for our **vayama deals newsletter** to receive our latest deals via email and automatically be entered into our monthly drawing to win a \$500 travel coupon.

Your picture here
Enter to win and this could be you!

Win a \$500 Travel Coupon official contest rules

\$30 OFF PER PERSON
USA TO SOUTH KOREA

GET DETAILS

ASIANA AIRLINES

Munich

Oktoberfest

\$30 Off
On Select

Commercial search engines are not built for educators.

Desired Search Criteria

- Text length
- Writing quality
- Target vocabulary
- Search by high-level topic
- Reading level

Familiar query box for specifying keywords.

REAP Search



Search for readings that match your pedagogical needs. [\[Help\]](#)

international travel

Target Words

consequence
maintain
evaluate
feature
emphasis
constant

Reading Grade Level: min max

Number of Words: min max

Topic:

Search

[Back to Teacher Menu](#)

Extra options for specifying pedagogical constraints.

User clicks **Search** and sees a list of results...

REAP Search Result

[Home](#) [Showcase Properties](#) [About Cebu](#) [Contact Us](#) [Sitemap](#)

Related Links

Areas of Interests

[History](#)
[Geography](#)
[People](#)
[Politics](#)
[Economy](#)
[Education](#)
[Languages](#)
[Transportation](#)
[Religion](#)
[Wildlife](#)

All About Travel

[Travel Destination](#)
[Dive Sites](#)
[Travel Tips](#)
[Travel News and Updates](#)

Cebu Real Estate

Property Listings

[Residential Properties](#)
[Commercial Properties](#)
[Agricultural Properties](#)
[Beach Front Properties](#)
[Residential Condominium](#)
[Bank Foreclosed Properties](#)
[Memorial Parks](#)

Philippine Travel News and Update

China Enormously Improves by 211.2% in Tourist Arrivals for May

Yet again, Peoples Republic of China sets an unprecedented record for the month of May as its registered number of visitors significantly escalated by 211.2%, or an increase to 9,806 warm bodies for 2005 from last years 3,151, based from the Department of Tourism (DOT) Statistics Division.

Furthermore, as early as mid-June, an exceptional upsurge of 482.56% has already been established by this tourism market, after gaining 4,276 tourists, a very huge step up from 2004s 734. This is obtained from the recent unofficial account from the NAIA daily visitor arrivals, which actually comprises 85% of the official figures.

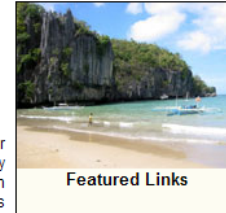
Tourism Secretary Joseph Ace Durano said, These results are indeed encouraging for us to further intensify our operation in China. Not only that, we are also very optimistic that we will be able to sustain this success with our current and future marketing campaigns.

Because China has shown noteworthy improvement over the past months, it has now become one of the countrys topmost sources of foreign travelers, joining the league of United States, Korea, and Japan. In fact, China has constantly rallied behind these countries, as far as number of visitors and growth rate are concerned.

In order to maintain this feat, the China marketing team of DOT headed by Tourism Assistant Secretary Eduardo Jarque, Jr. and members Director Rolando Caizal, Rene Reyes, Mila Say, and Gigi Liwanag has pushed for priority policy support here in the country. The Meet and Assist program, which caters to non-English-speaking guests like the Chinese, Koreans and Japanese, is being instigated now at international airports in both Manila and Cebu. Still particularly designed for Chinese visitors, the innovative Visa-Upon-Arrival program has been operational through the cooperation of Department of Foreign Affairs and the Bureau of Immigration and helps tourists in processing their visa. Moreover, there are ongoing trainings for Mandarin-speaking guides to address possible language barrier problems.

In addition to the DOT office in Beijing, a marketing representative based in Shanghai was appointed last July 15.

The Philippines has also attended different travel fairs and consumer

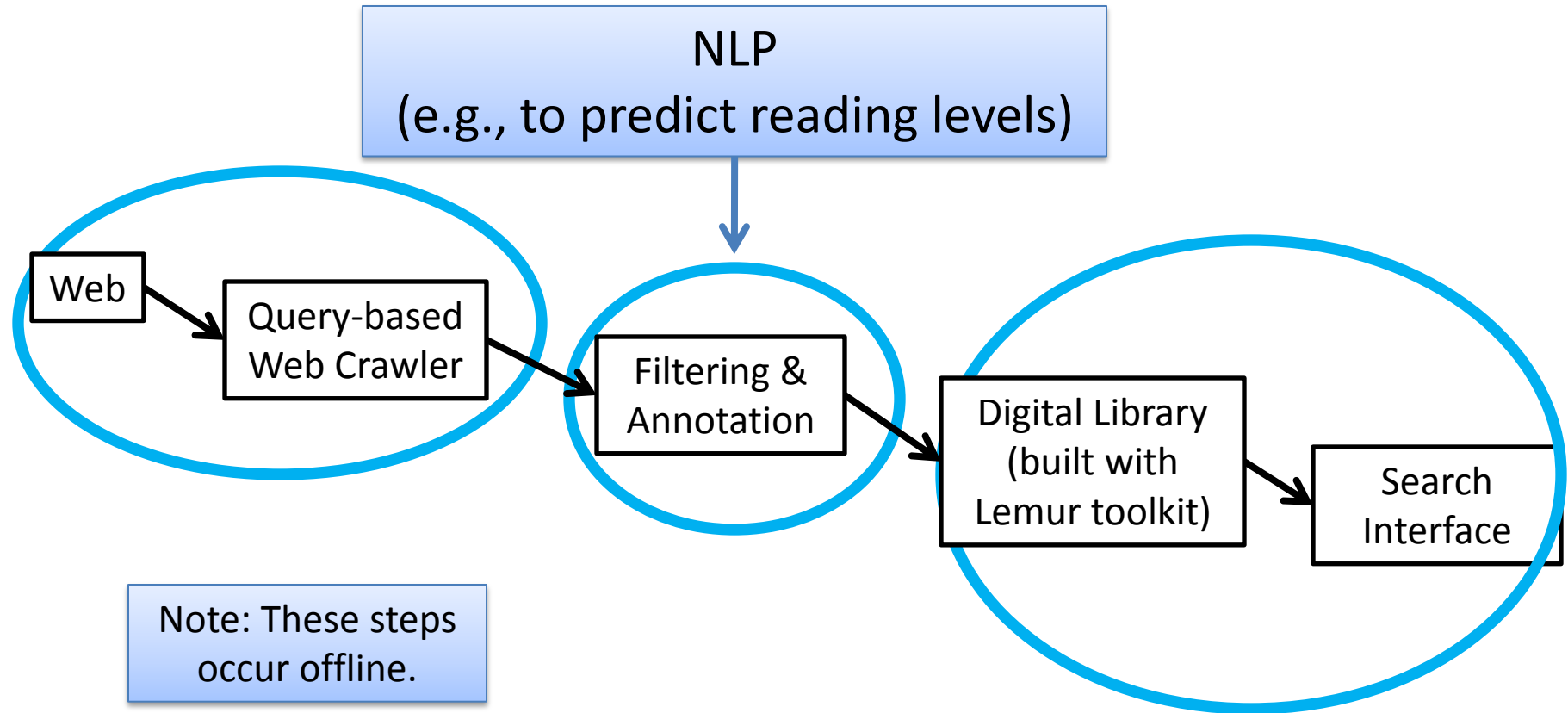


Featured Links

Outline

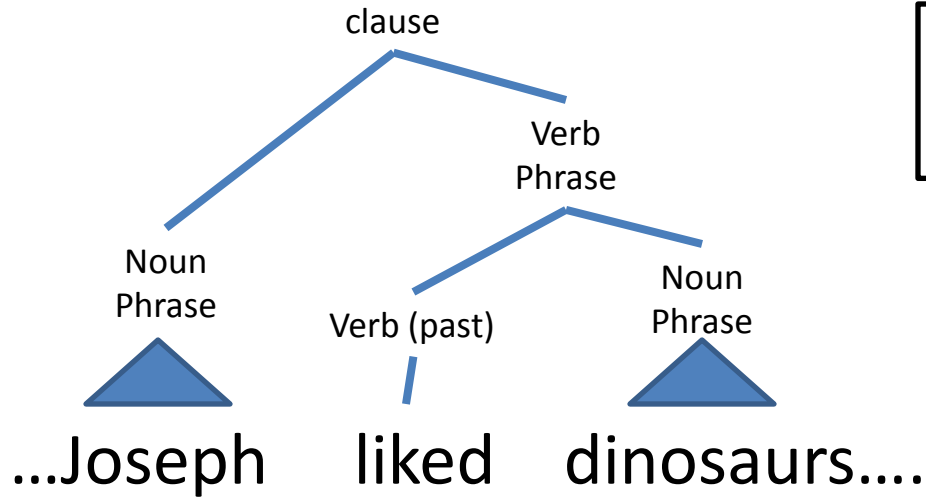
- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
 - Motivation
 - NLP components
 - Pilot study
- Question Generation
- Concluding Remarks

Digital Library Creation



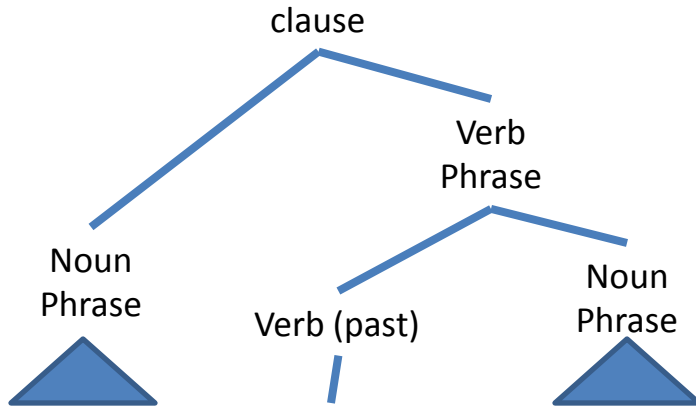
Heilman, Zhao, Pino, and Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. 3rd Workshop on NLP for Building Educational Applications.

Predicting Reading Levels



Simple syntactic structure
==> **low** reading level

Predicting Reading Levels



...Thoreau apotheosized nature....

Simple syntactic structure
==> **low** reading level

Infrequent lexical items
==> **high** reading level

We can use statistical NLP techniques to estimate weights from data.

We need to adapt NLP for specific tasks.
(e.g., to specify important linguistic features)

Potentially Useful Features for Predicting Reading Levels

- Number of words per sentence
- Number of syllables per word
- Depth/complexity of syntactic structures
- Specific vocabulary words
- Specific syntactic structures
- Discourse structures
- ...

For speed and scalability,
we used a vocabulary-based
approach
(Collins-Thompson & Callan, 05)

Flesch-Kincaid, 75; Stenner et al. 88; Collins-Thompson & Callan, 05; Schwarm & Ostendorf 05; Heilman et al., 07;
inter alia

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
 - Motivation
 - NLP components
 - Pilot study
- Question Generation
- Concluding Remarks

Pilot Study

Participants

- 2 instructors and 50+ students
- Pittsburgh Science of Learning Center's English LearnLab
- Univ. of Pittsburgh's English Language Institute

Typical Usage

- Before class, teachers found texts using the tool
- Students read texts individually
- Also, the teachers led group discussions
- 8 weeks, 1 session per week

Evidence of Student Learning

- Students scored approximately 90% on a post-test on target vocabulary words
- Students also studied the words in class.
- There was no comparison condition.

More research is needed

Teacher's Queries

$$\frac{47 \text{ unique queries}}{23 \text{ selected texts used in courses}} = 2.04 \text{ queries to find a useful text (on average)}$$

The digital library contained
3,000,000 texts.

Teacher's Queries

Teachers found high-quality texts, but often had to relax their constraints.

Exaggerated Example:

- 7th grade reading-level
- 600-800 words long
- 9+ vocabulary words from curriculum
- keywords: “construction of Panama Canal”



- 6-9th grade reading-level
- less than 1,000 words long
- 3+ vocabulary words
- topic: history

Teacher's Queries

Teachers found high-quality texts, but often had to relax their constraints.

Possible future work:

- Improving the accuracy of the NLP components
- Scaling up the digital library

Related Work

System	Reference	Description
REAP Tutor	Brown & Eskenazi, 04	A computer tutor that selects texts for students based on their vocabulary needs (also, the basis for REAP search).
WERTi	Amaral, Metcalf, & Meurers, 06	An intelligent automatic workbook that uses Web texts to teach English grammar.
SourceFinder	Sheehan, Kostin, & Futagi, 07	An authoring tool for finding suitable texts for standardized test items.
READ-X	Miltsakaki & Troutt, 07	A tool for finding texts at specified reading levels.

REAP Search...

- Applies various NLP and text retrieval technologies.
- Enables teachers to find pedagogically appropriate texts from the Web.

For more recent developments in the REAP project, see <http://reap.cs.cmu.edu>.

Segue

- So, we can find high quality texts.
- We still need exercises and assessments...

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- **Question Generation**
- Concluding Remarks

Question Generation
Collaborator:
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

Our Approach

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

Heilman and Smith. 2010. Good Question! Statistical Ranking for Question Generation. In Proc. of NAACL/HLT.

Outline

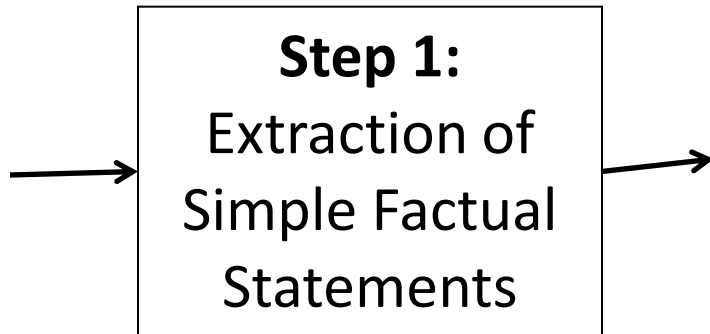
- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation
 - Challenges
 - Step-by-step example
 - Question ranking
 - User interface
- Concluding Remarks

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?

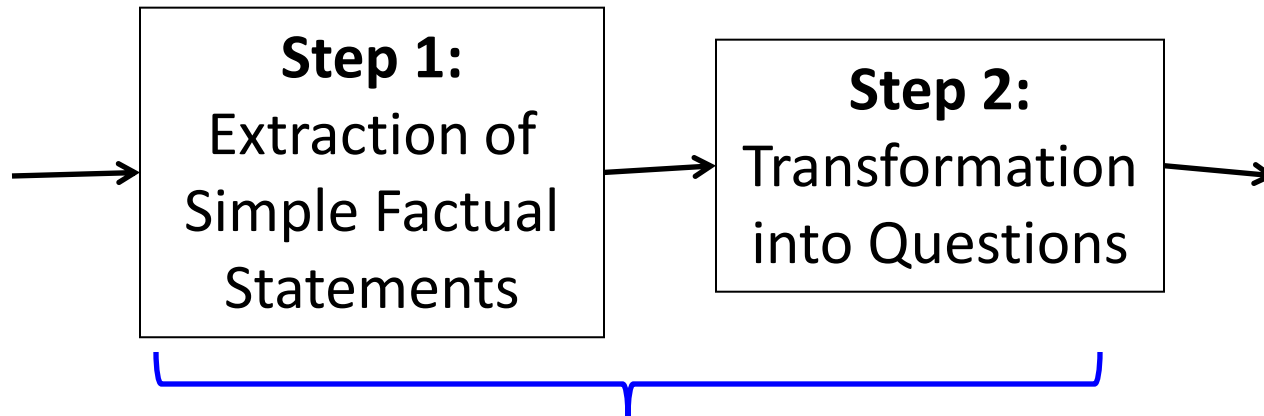


Constraints on Question Formation

Darwin studied how species evolve.

Who studied how species evolve? 😊

*What did Darwin study how evolve? 😞



Rules that encode linguistic knowledge

Vague and Awkward Questions, etc.

Lincoln, who was born in Kentucky...

Where was Lincoln born?



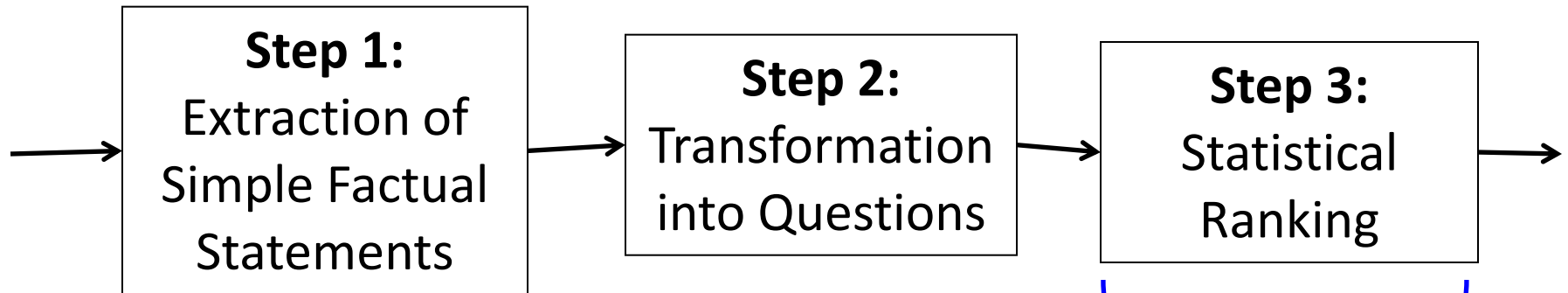
Lincoln, who faced many challenges...

What did Lincoln face?



Weak predictors:

proper nouns,
who/what/where...,
sentence length,
etc.



Model learned from human-rated
output from steps 1&2

Step 0: Preprocessing with NLP Tools

- Stanford parser Klein & Manning, 03
 - To convert sentences into syntactic trees
- Supersense tagger Ciaramita & Altun, 06
 - To label words with high level semantic classes (e.g., person, location, time, etc.)
- Coreference resolver <http://www.ark.cs.cmu.edu/arkref>
 - To figure out what pronouns refer to

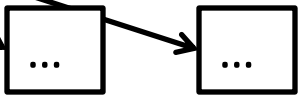
Each may introduce errors that lead to bad questions.

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation
 - Challenges
 - Step-by-step example
 - Question ranking
 - User interface
- Concluding Remarks



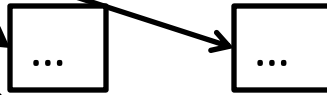
(other candidates)



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.

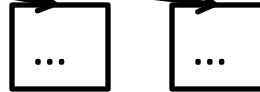
Extraction of Simplified Factual Statements



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. ...

Existing Work on QG

Reference	Description
Wolfe, 1977	Early work on the topic.
Mitkov & Ha, 2005	Template-based approach based on surface patterns in text.
Heilman & Smith, 2010	Over-generation and statistical ranking.
Mannem, Prasad, & Joshi, 2010	QG from semantic role labeling analyses.
<i>inter alia</i>	

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation
 - Challenges
 - Step-by-step example
 - Question ranking
 - User interface
- Concluding Remarks

Question Ranking

We use a statistical ranking model to avoid vague and awkward questions.

Logistic Regression of Question Quality

$$y \in \{ \text{😊}, \text{😞} \}$$

x : features of the question
(binary or real-valued)

w : weights
(learned from labeled questions)

$$\log P(y = \text{😊}) \propto \vec{x} \cdot \vec{w}$$

To rank, we sort by $\log P(y = \text{😊})$

Surface Features

- Question words (who, what, where...)
 - e.g., $x_j = 1.0$ if “What...”
- Negation words
- Sentence lengths
- Language model probabilities
 - a standard feature to measure fluency

Features based on Syntactic Analysis

- Grammatical categories
 - Counts of parts of speech, etc.
 - e.g., if 3 proper nouns, $x_j = 3.0$
- Transformations
 - e.g., extracted from relative clause
- “Vague noun phrase”
 - distinguishes phrases like “the president” from “Abraham Lincoln” or “the U.S. president during the Civil War”

Feature weights

- We estimate weights from a training dataset of human-labeled output from steps 1 & 2.

Feature (x_j)	Weight (w_j)
Question starts with “when”	0.323
Past tense	0.103
Number of proper nouns	0.052
Negation words in the question	-0.144
...	...

Evaluation

- We generated questions about texts from Wikipedia and the Wall Street Journal.
- Human judges rated the output.
- **27%** of **unranked** questions were acceptable.
- **52%** of the **top-ranked** fifth were acceptable.

Heilman and Smith, 2010.

System Output

(from a text about Copenhagen)

What is the home of the Royal Academy of Fine Arts?

(Answer: the 17th-century Charlottenborg Palace)

Who is the largest open-air square in Copenhagen?

(Answer: Kongens Nytorv, or King's New Square)

About one third of bad questions result from preprocessing errors.

What is also an important part of the economy?

(Answer: ocean-going trade)

The system still makes many errors.

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation
 - Challenges
 - Step-by-step example
 - Question ranking
 - User interface
- Conclusion

source text

ranked question candidates

A Tool for Generating Factual Questions

The screenshot shows a web interface with three main columns. The left column, titled 'Article', contains text about Ethiopia's government and Addis Ababa, with a yellow highlight on the sentence 'Addis Ababa University was established in 1950.' The middle column, titled 'Question', lists three ranked questions: 'What was established in 1950?', 'Was Addis Ababa University established in 1950?', and 'When was Addis Ababa University established?'. The second question is highlighted in yellow, and an 'Answer: in 1950' box is shown below it. The right column, titled 'Quick Search', lists shortcuts: 'who', 'what', 'where', 'when', 'beginning', 'middle', 'end', and 'all questions'. At the bottom, there is a 'Selected Questions' section with two rows. The first row shows a question 'Who paved roads and constructed European-style buildings?' with the answer 'the Italians' and a 'delete' button. The second row shows 'When was Addis Ababa University established?' with the answer 'in 1950' and a 'delete' button. A 'keyword search box' with a 'Search' button is located below the 'Question' column. A 'text version' link and an 'add your own question' link are also present.

shortcuts

keyword search box

user-selected questions (editable)

option to add your own question

User Feedback

- Adding one's own questions is important
 - “Deeper” questions
 - Reading strategy questions
- Easy-to-use interface
- Differing opinions about specific features
 - e.g., search, document-level vs. sentence-level
- Shareable questions

Outline

- Introduction
- Textbooks vs. New Resources
- Text Search for Language Instructors
- Question Generation
- Concluding Remarks

NLP is not a black box

- NLP must be adapted for specific applications.
 - Labeled data and linguistic knowledge are often needed.
 - Of course, applications for other languages are possible....



- One must consider how to handle errors.

An Analogy: Chinese food in America

- Good
- Fast
- Cheap

You pick 2









An Analogy: Natural Language Processing

- **high accuracy**
- **broad domain (not just for a single topic)**
- fully automatic

Educators need to check the output.

Some Example Applications

	Google Translate	Phone systems (e.g., for banking)	This research
high accuracy			
broad domain			
fully automatic			

Summary

- Vast resources of text are available.
- We can develop NLP tools to help teachers use those resources.
 - NLP is not magic (e.g., we need to handle errors).
- Specific applications:
 - Search tool for reading materials
 - Factual question generation tool

Question Generation demo: <http://www.ark.cs.cmu.edu/mheilman/questions>

References

- M. Heilman, L. Zhao, J. Pino, and M. Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In Proc. of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications.
- M. Heilman and N. A. Smith. 2010. Good Question! Statistical Ranking for Question Generation. In Proc. of NAACL/HLT.
- M. Heilman, A. Juffs, and M. Eskenazi. 2007. Choosing reading passages for vocabulary learning by topic to increase intrinsic motivation. In Proc. of AIED.
- K. Collins-Thompson and J. Callan. 2005. Predicting reading difficulty with statistical reading models. Journal of the American Society for Information Science and Technology.

Prior Work on Readability

Measure	Year	Lexical Features	Grammatical Features
Flesch-Kincaid	1975	Syllables per word	Sentence length
Lexile (Stenner, et al.)	1988	Word frequency	Sentence length
Collins-Thompson & Callan	2004	Individual words	-
Schwarm & Ostendorf	2005	Individual words & sequences of words	Sentence length, distribution of POS, parse tree depth, ...
Heilman, Collins-Thompson, & Eskenazi	2008	Individual words	Syntactic sub-tree features

Curriculum Management Interface

Enables teachers to...

- Search for texts,
- Order presentation of texts,
- Set time limits,
- Choose vocabulary to highlight,
- Add practice questions.

Learner Support: Reading Interface

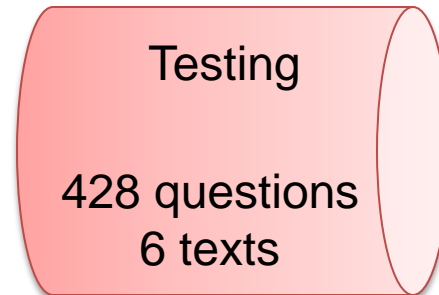
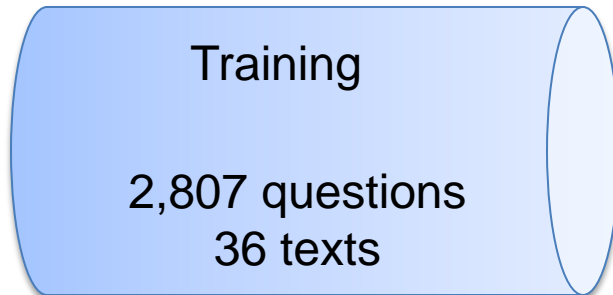
The screenshot shows a Mozilla Firefox browser window displaying a reading page titled "Global Warming : Climate Change". The page content includes the text: "Most everyone has heard about climate change, and most certainly about **global** warming. While some people still don't believe climate change exists, I am convinced it is real." and "In January, climate **experts** from 30 countries met in England to discuss new evidence which proves we are fast approaching the point of no return. These **experts** report an ecological time bomb ticking away toward widespread drought, crop failures and rising sea levels. Scientists throughout the world continue to conclude with deep urgency that climate change is creating dangerous conditions that **require** immediate attention."

Callouts and features shown:

- Target words specified by the teacher are highlighted.** (Points to the word "experts" in red in the text.)
- Students click on target words for definitions.** (Points to a definition pop-up window for "expert (n)" which includes: "a person with a high level of knowledge or skill, a specialist" and "Examples: a gardening/medical expert, My mother is an expert [at] dress-making (she does it very well).")
- Definitions available for non-target words as well.** (Points to a search box containing "expert" and a "Look up a word" button.)
- Optional timer helps with classroom management.** (Points to a "12 minutes left" timer and a "Done reading -->" button.)

Corpora

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



Evaluation Metric

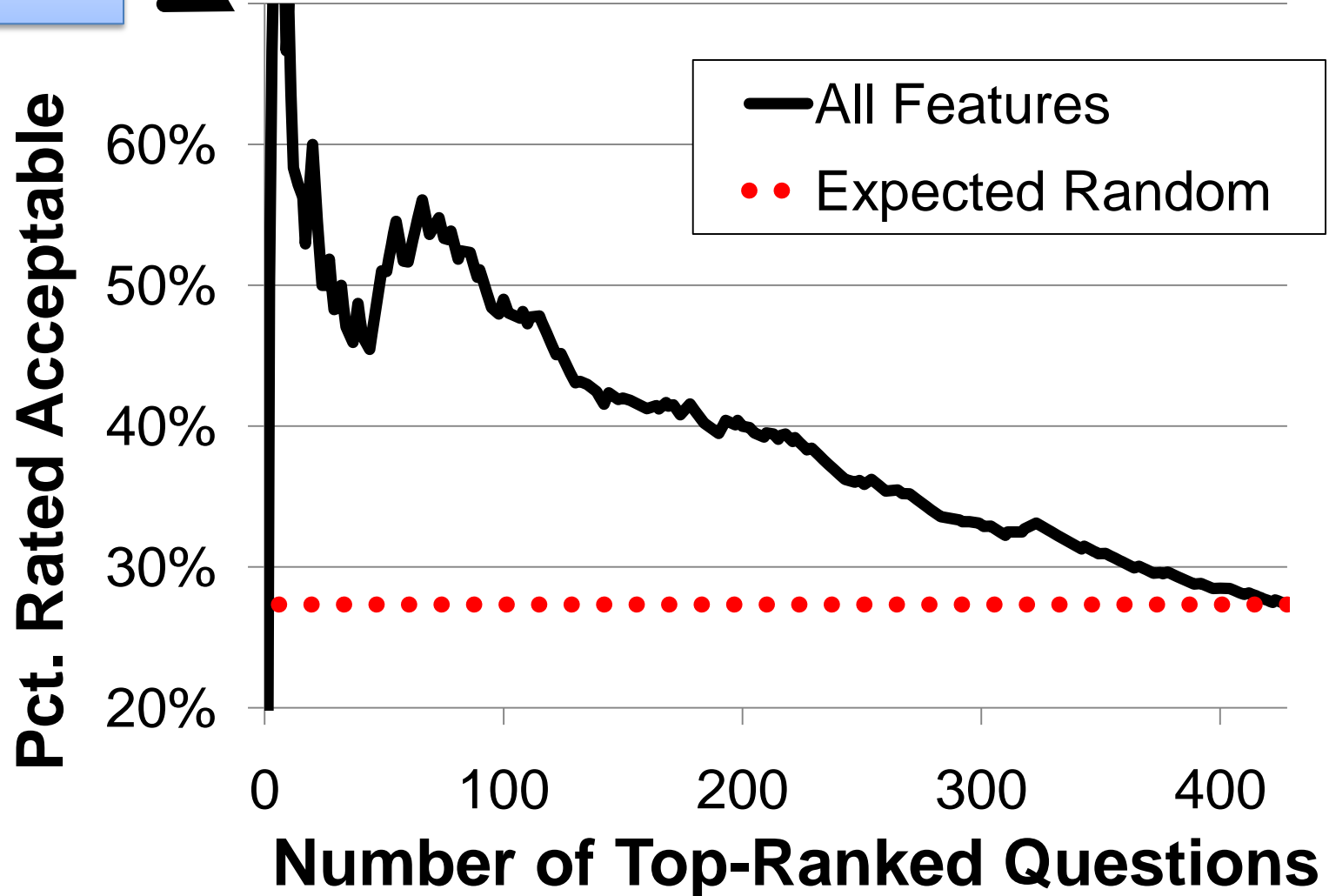
Percentage of top-ranked test set questions that were rated acceptable by human annotators



Ranking Results

Testing

Noisy at top ranks.



Selecting and Revising Questions

(location) (person) (location)
...Jefferson, the third President of the U.S.,
(person) (person)
selected Aaron Burr as his Vice President....

revision
by a user



Where was the third President of the U.S.?
Who was the third President of the U.S.?