Teaching the Web

Learning How to Use Arbitrary Text as Teaching Material

Sumit Basu and Lucy Vanderwende
(and star intern Lee Becker on the question generation work)

A Few Scenarios

Mary
home buyer

Jae
interview prep

Karthik
diabetes management

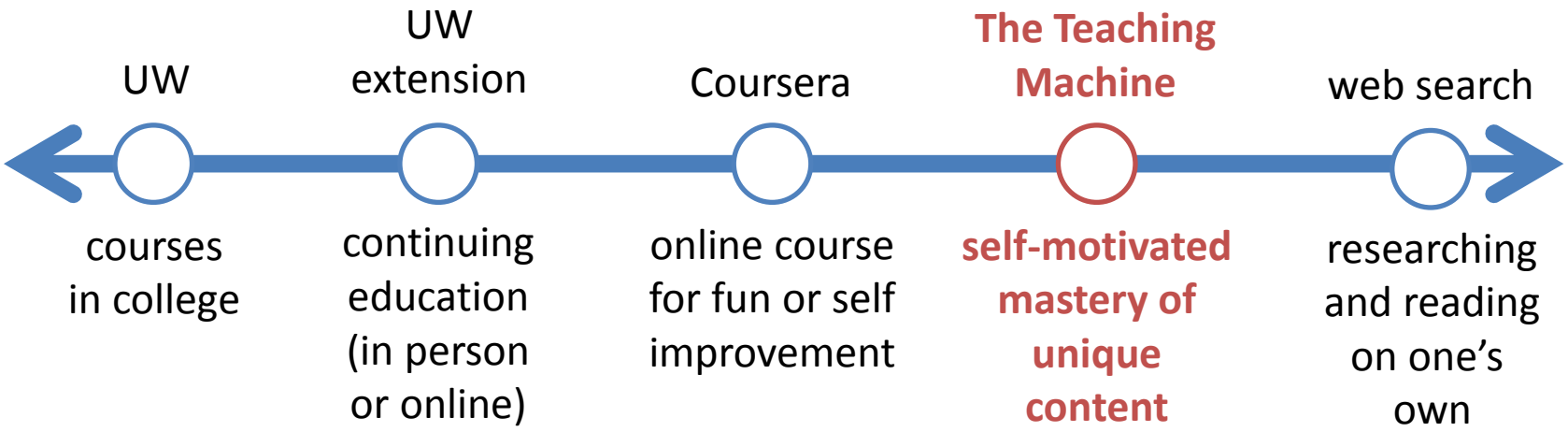
Nina
leadership

Common Threads:

-Self-motivated learners
-Wide variety of sources
-Factual and conceptual material
-A need for mastery

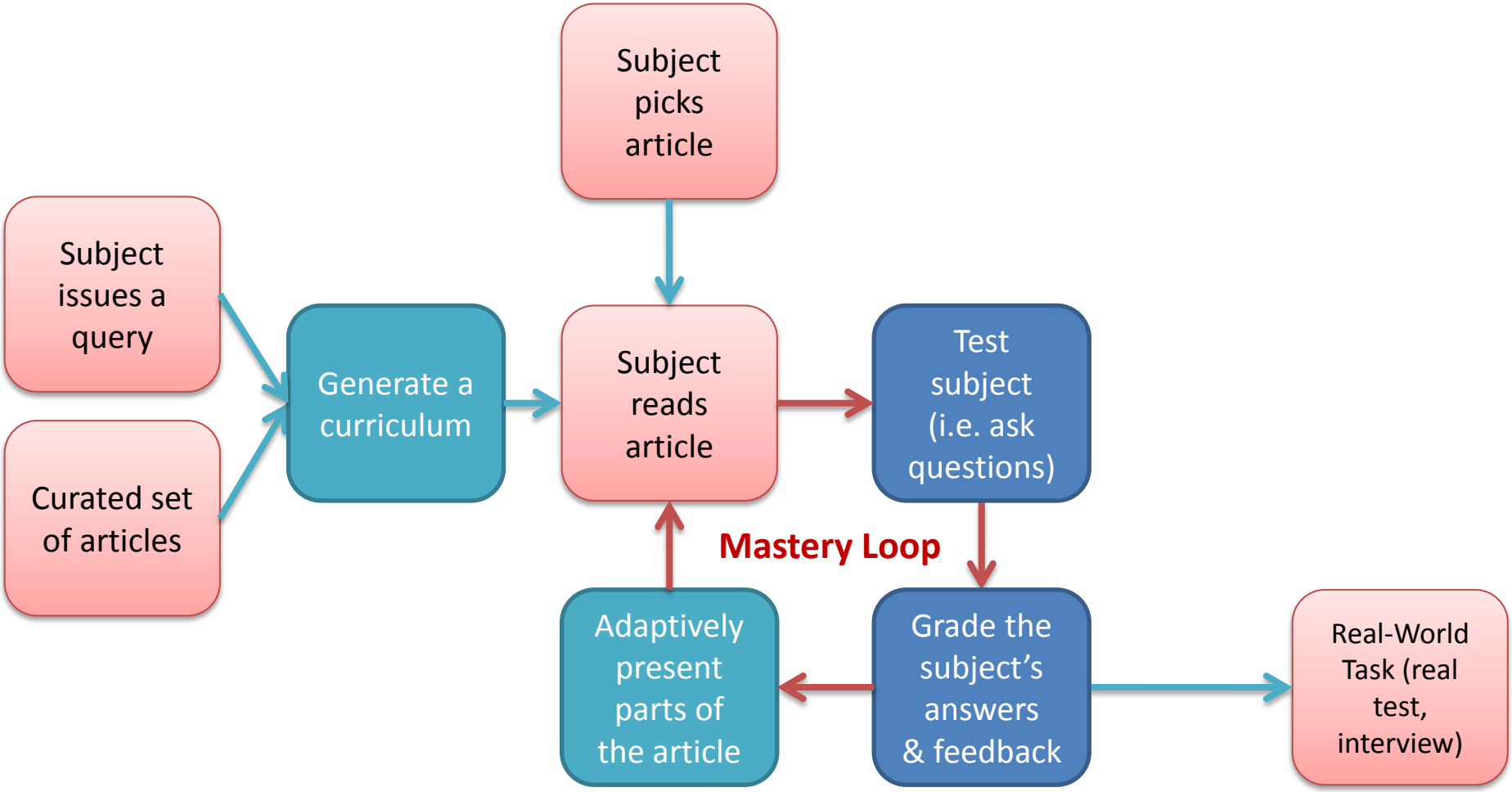
Raven
web services

Where This Type of Learning Fits In



The Formality Spectrum of Educational Experiences

Vision of "The Teaching Machine"



The Value of Testing

The Critical Importance of Retrieval for Learning

Jeffrey D. Karpicke[1*] and Henry L. Roediger III[2]

Learning is often considered complete when a student can produce the correct answer to a question. In our research, students in one condition learned foreign language vocabulary words in the standard paradigm of repeated study-test trials. In three other conditions, once a student had correctly produced the vocabulary item, it was repeatedly studied but dropped from further testing, repeatedly tested but dropped from further study, or dropped from both study and test. Repeated studying after learning had no effect on delayed recall, but repeated testing produced a large positive effect. In addition, students' predictions of their performance were uncorrelated with actual performance. The results demonstrate the critical role of retrieval practice in consolidating learning and show that even university students seem unaware of this fact.

Ever since the pioneering work of Ebbinghaus (*1*), scientists have generally studied human learning and memory by presenting people with information to be learned in a study period and testing them on it in a test period to see what they retained. When this procedure occurs over many trials, an exponential learning curve is produced. The standard assumption in nearly all research is that learning occurs while people study and encode material. Therefore, additional study should increase learning. Retrieving information on a test, however, is sometimes considered a relatively neutral event that measures the learning that occurred during study but does not by itself produce learning. Over the years, researchers have occasionally argued that learning can occur during testing (*2–6*). However, the assumptions that repeated studying promotes learning and that testing represents a neutral event that merely measures learning still permeate contemporary memory research as well as contemporary educational practice, where tests are also considered purely as assessments of knowledge.

Our goal in the present research was to examine these long-standing assumptions regard-

memory, concerning the relation between the speed with which something is learned and the rate at which it is forgotten. Is speed of learning correlated with long-term retention, and if so, is the correlation positive (processes that promote fast learning also slow forgetting and promote good retention) or negative (quick learning may be superficial and produce rapid forgetting)? Early research led to the conclusion that quick learning reduced the rate of forgetting and improved long-term retention (*7*), but later critics argued that, when forgetting is assessed more properly than in the early studies, no differences exist between forgetting rates for fast and slow learning conditions (*8*, *9*). By any account, conditions that exhibit equivalent learning curves should produce equivalent retention after a delay (*9*).

Using foreign language vocabulary word pairs, we examined the contributions of repeated study and repeated testing to learning by comparing a standard learning condition to three dropout conditions. The standard method of measuring learning, used since Ebbinghaus's research (*1*), involves presenting subjects with information in a study

- **Karpicke and Roediger**, 2008, "The Critical Importance of Retrieval for Learning."
- **Anderson and Biddle**, 1975, "On Asking People Questions About What They are Reading."
- **Laufer and Goldstein**, 2004, on the difficulty of Recall tasks vs. Recognition
- **The Dunning-Kruger effect**: the cognitive bias in which the unskilled think they have mastery
- **McGraw-Hill representatives** – the persistent need for new tests for teachers (helper tool) and students (self-review)

i.e., Teaching Requires Assessment

- Mastery is achieved through repeated presentation and testing
- We want to teach using web materials
  - Our (first) goal: generate (cloze) questions from arbitrary web material.
  - Specifically: train a ML model to target **the best spans of the best sentences** to ask questions about.

Question Generation Overview

1.

Sentence Selection

2. Like Pierre Curie, Röntgen refused to take out patents related to his discovery.

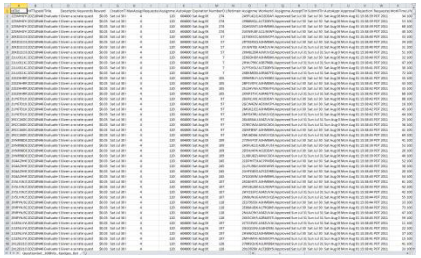
Candidate Construction

Like _____, Röntgen refused to take out patents related to his discovery.
Like Pierre Curie, Röntgen refused to take out _____ related to his discovery.

3.

Question	Answer	
The _ _ _ _ _ _ _ _ of chemicals was an important development during the Industrial Revolution.	large scale production of chemicals	Good Okay Bad
The large scale production of _ _ _ _ was an important development during the Industrial Revolution.	chemicals	Good Okay Bad
The large scale production of chemicals was an important development during the _ _ _ _ _ _ _ _.	Industrial Revolution	Good Okay Bad

Corpus Construction



4. Like _____, Röntgen refused to take out patents related to his discovery.
Like Pierre Curie, Röntgen refused to take out _____ related to his discovery.

Candidate Scoring

Like _____, Röntgen refused to take out patents related to his discovery. 0.6
Like Pierre Curie, Röntgen refused to take out _____ related to his discovery. 0.8

Lee Becker, Sumit Basu, and Lucy Vanderwende. "Mind the Gap: Learning to Choose Gaps for Question Generation." *NAACL 2012*.

Related Work

- Wh-Questions
  - Autoquest (Wolfe, 1976)
  - Transformation rules (Mitkov and Ha, 2003)
  - Template-based generation (Chen et al., 2009)
  - Overgenerate-and-rank (Heilman and Smith, 2010)
  - QG-STEC (Rus et al., 2010)
- Fill-in-the-blank (aka gap-fill & cloze) questions
  - Content-focused, heuristic (Agarwal and Mannem, 2011)
  - Vocabulary and language learning (Pino et al., 2008)

Sentence Selection

- **Intuition**
  - Parallels to extractive summarization
  - Key material first, obscure material later
- **Approach**
  - The SumBasic Algorithm [Nenkova et al., 2006]

Candidate Question Construction and Scoring

- **Task**: Given a sentence, generate a question that best covers the material in that sentence.
- **Metric**: Human judgment (via crowdsourcing)
- **Approach**: Overgenerate and rank
  - Identify candidate blanks
  - Extract context independent features
  - Train/Evaluate ‘Good’ vs ‘not-Good’ question classifier.

Gap Generation Example

- Before Genghis Khan died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

Gap Generation Example

- Before Genghis Khan died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

1. Before _____died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

2. Before Genghis Khan ____, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

Semantic Role Labels:	**Pred**	**A0**	**A1**	**A2**	**AM-TMP**

Gap Generation Example

- Before Genghis Khan died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

1. Before _____ died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.
2. Before Genghis Khan died, __ assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.
3. Before Genghis Khan died, he _____ Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.
4. Before Genghis Khan died, he assigned _____ as his successor and split his empire into khanates among his sons and grandsons.
5. Before Ghengis Khan died, he assigned Ögedei Khan as _____ and split his empire into khanates among his sons and grandsons.

Semantic Role Labels:	**Pred**	**A0**	**A1**	**A2**	**AM-TMP**

Gap Generation Example

- Before Genghis Khan died, he assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.

1. Before Genghis Khan died, __ assigned Ögedei Khan as his successor and split his empire into khanates among his sons and grandsons.
2. Before Genghis Khan died, he assigned Ögedei Khan as his successor and ____ his empire into khanates among his sons and grandsons.
3. Before Genghis Khan died, he assigned Ögedei Khan as his successor and split _____ into khanates among his sons and grandsons.
4. Before Genghis Khan died, he assigned Ögedei Khan as his successor and split his empire into _____.

Semantic Role Labels: **Pred** **A0** **A1** **A2** **AM-TMP**

Crowdsourcing a Corpus

- Downloaded 105 vital/popular Wikipedia articles
- Sentences:
  - SumBasic: 10 per article
  - Random Sampling: 10 per article
- Generate candidate questions
- Collect ratings of question quality via Amazon Mechanical Turk
- Full corpus available at http://research.microsoft.com/~sumitb/questiongeneration

Crowdsourcing QG: HITs

"The large scale production of chemicals was an important development during the Industrial Revolution."

Question	Answer	
The _ _ _ _ _ _ _ _ _ _ of chemicals was an important development during the Industrial Revolution.	large scale production of chemicals	○ Good ○ Okay ○ Bad
The large scale production of _ _ _ _ was an important development during the Industrial Revolution.	chemicals	○ Good ○ Okay ○ Bad
The large scale production of chemicals was an important development during the _ _ _ _ _ _ _ _ _ _.	Industrial Revolution	○ Good ○ Okay ○ Bad

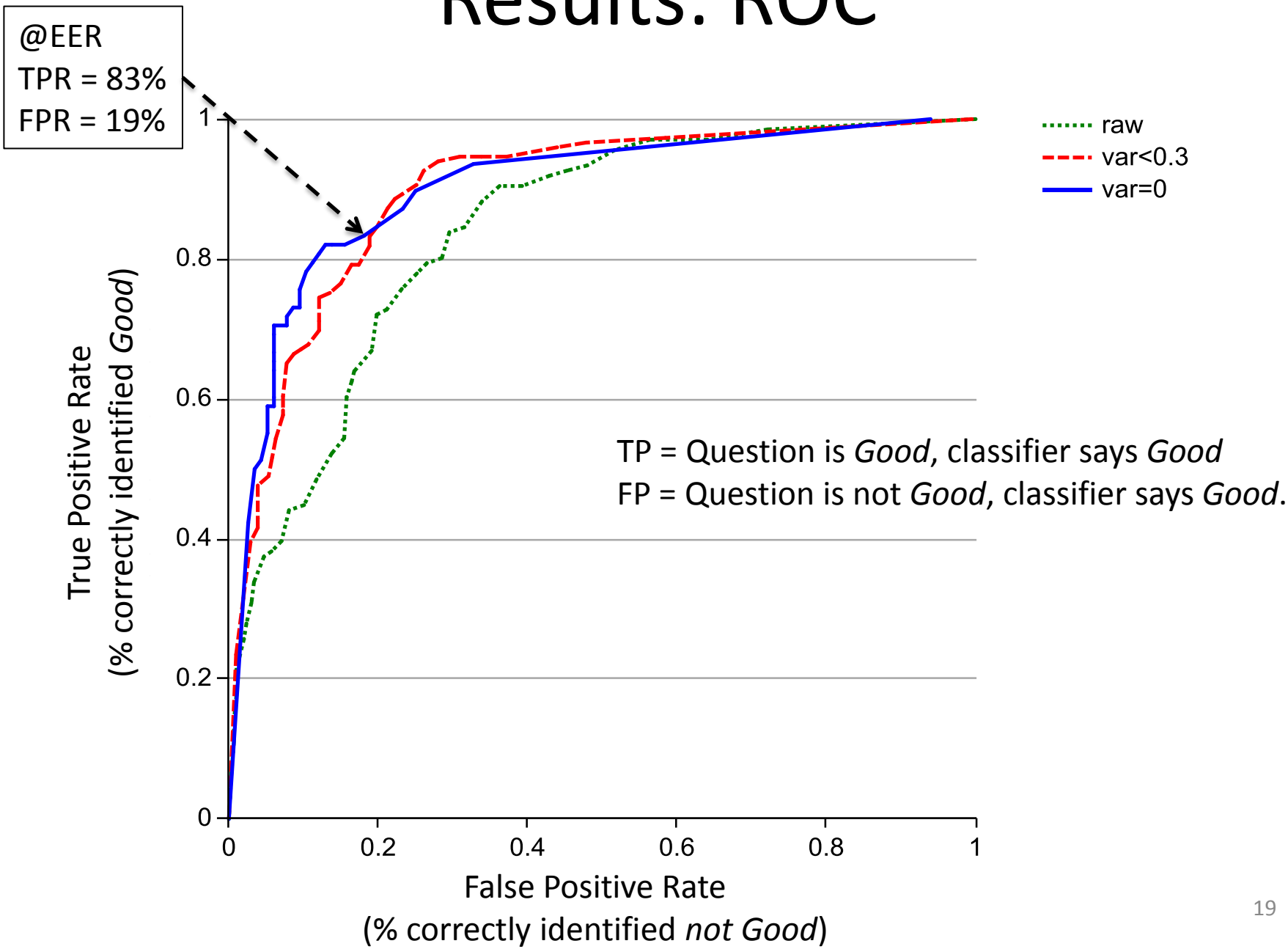
Filtering Turkers and Questions

- **Turkers:** Computed median judgment and distances to median
  - Removed judges with a mean distance two-standard deviations above the mean distance (5 judges)
- **Questions:** Computed variance of judgments for each question:
  - Limiting variance to 0.3, keeps disagreement to 1 out of 4 judges
  - Eliminated 431 questions, Retained 1821 (out of 2252) questions with highest agreement.
  - Of filtered questions 700 (38%) labeled *Good*

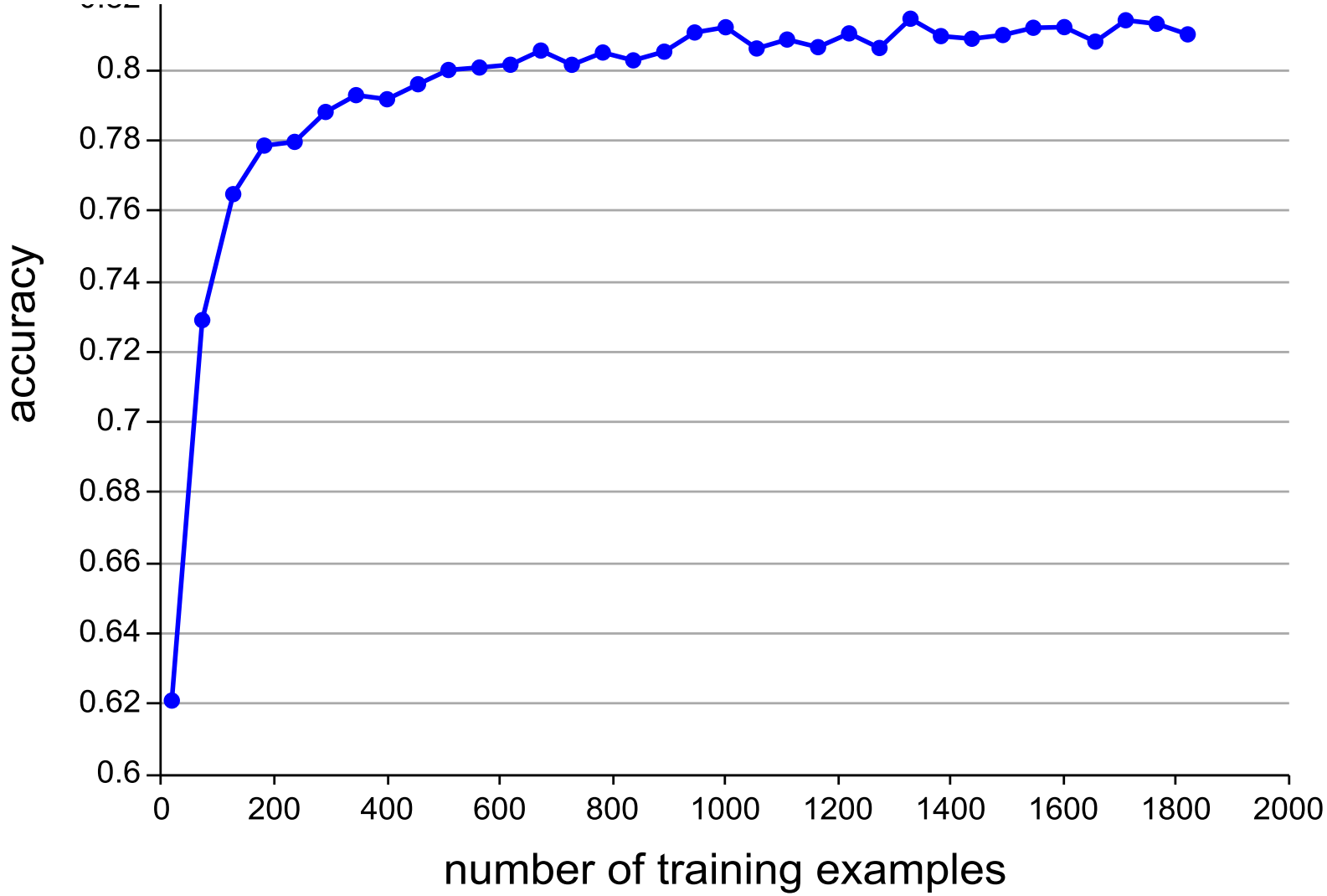
Learning a Model for Gap Selection

- Approach: Overgenerate and score:
  - Identify candidate blanks
  - Extract features from the sentence and the gap
  - Train/Evaluate ‘Good’ vs ‘not-Good’ question classifier.
  - For scoring use calibrated learner
    - Logistic Regression + L2 Regularizer
  - Evaluation: 10-fold cross validation

Results: ROC

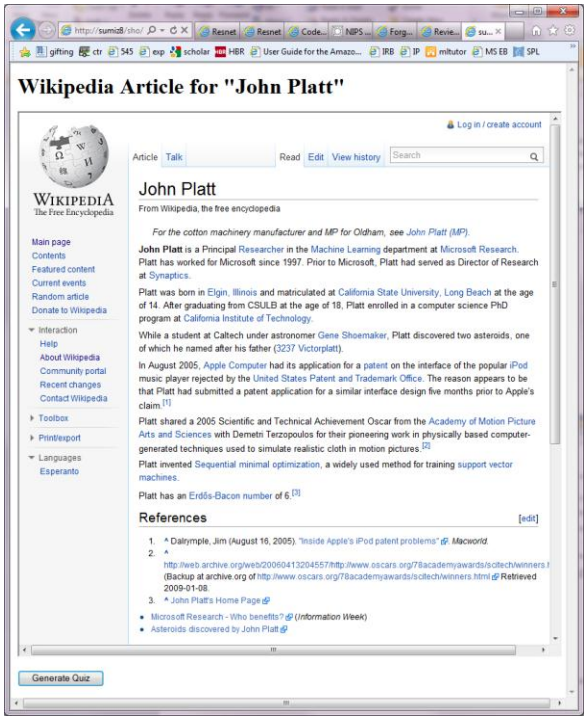

@EER
TPR = 83%
FPR = 19%
raw
var<0.3
var=0
TP = Question is *Good*, classifier says *Good*
FP = Question is not *Good*, classifier says *Good*.
True Positive Rate
(% correctly identified *Good*)
False Positive Rate
(% correctly identified *not Good*)
0
0.2
0.4
0.6
0.8
1

Learning Curve

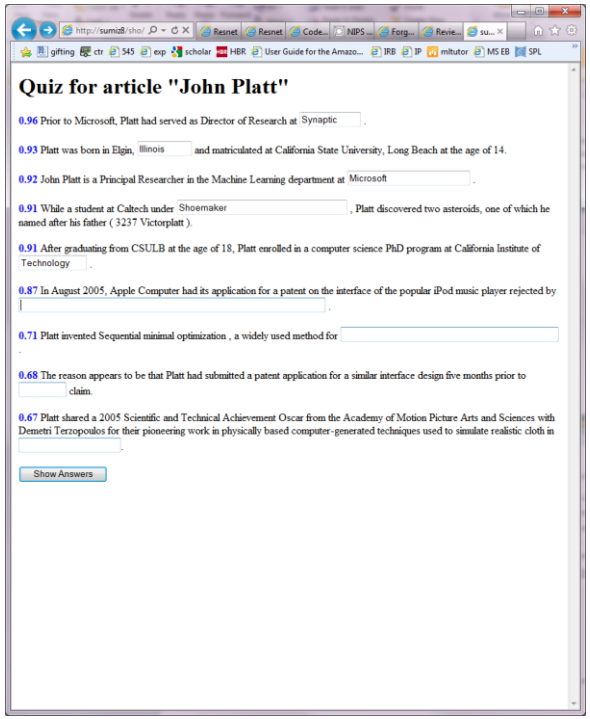


Question Generation Demo

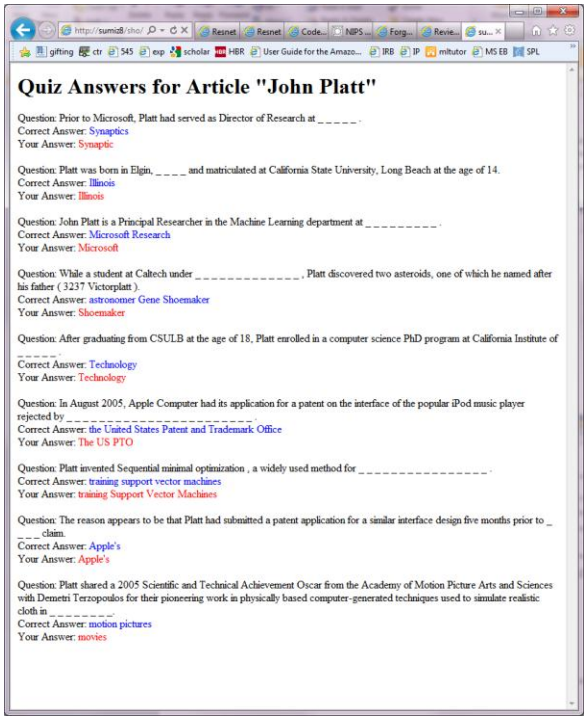
Read the article…



Fill in the blanks…



Check your answers!



Now: Grading Questions

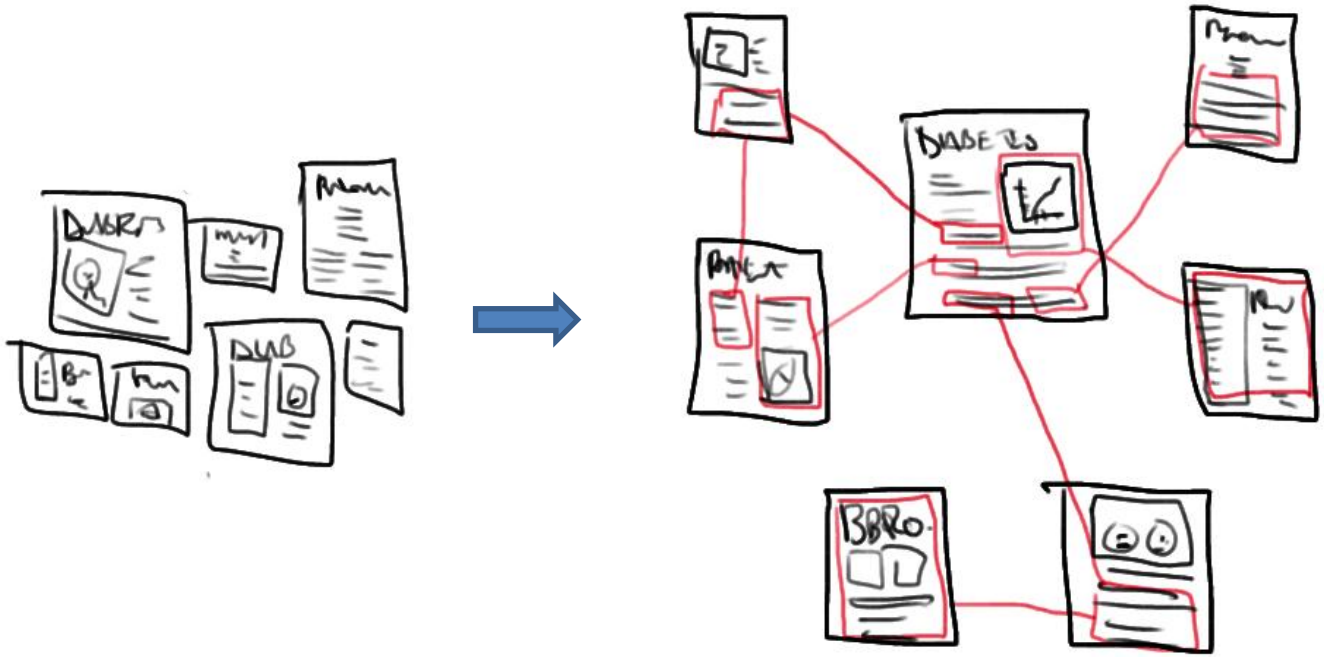
- How can we grade fill-in-the-blank questions?
- Can we do it quickly, cheaply, accurately?
- Gave 1280 sections to Turkers (320x4 judges), 5 q's each (6400 total)
  - 1: turkers read section
  - 2: we hid the section and gave them the quiz
  - 3: they saw the true answer and their own, asked to self-grade
- 984 items graded by two experts (Sumit/Lucy)
- 911 items where experts gave the same grade
- We also distributed first 1000 questions to other Turkers to grade
- Next step – a calibrated automatic means of grading that can shunt to Turkers

Table 1: Agreement of various methods with experts on the 911 question/answer pairs where both experts agreed on the grade

Method	Agreement	More Harsh	More Lenient
Self Grading	93.5%	4.5%	2.0%
Turker Grading	95.4%	2.4%	2.2%
String Match	79.1%	20.9%	0.0%

Next: Multiple Documents

query



Goal: understand the relationship between multiple documents to create a guided, filterable, interactive reading experience

Discussion and Questions