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

Common Threads:
- Self-motivated learners
- Wide variety of sources
- Factual and conceptual material
- A need for mastery
Where This Type of Learning Fits In

The Formality Spectrum of Educational Experiences

- UW courses in college
- UW extension continuing education (in person or online)
- Coursera online course for fun or self improvement
- The Teaching Machine self-motivated mastery of unique content
- web search researching and reading on one’s own
Vision of “The Teaching Machine”

- Subject issues a query
- Curated set of articles
- Generate a curriculum
- Subject picks article
- Subject reads article
- Test subject (i.e. ask questions)
- Grade the subject’s answers & feedback
- Adaptively present parts of the article
- Mastery Loop
- Real-World Task (real test, interview)
The Value of Testing

- **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.

3. **Candidate Construction**

4. Like ________, Röntgen refused to take out patents related to his discovery.

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/question generation
Crowdsourcing QG: HITs

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

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>The ___________ of chemicals was an important development during the Industrial Revolution.</td>
<td>large scale production of chemicals</td>
</tr>
<tr>
<td>The large scale production of __________ was an important development during the Industrial Revolution.</td>
<td>chemicals</td>
</tr>
<tr>
<td>The large scale production of chemicals was an important development during the __________.</td>
<td>Industrial Revolution</td>
</tr>
</tbody>
</table>
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%

TP = Question is *Good*, classifier says *Good*
FP = Question is not *Good*, classifier says *Good*.
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Agreement</th>
<th>More Harsh</th>
<th>More Lenient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self Grading</td>
<td>93.5%</td>
<td>4.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Turker Grading</td>
<td>95.4%</td>
<td>2.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td>String Match</td>
<td>79.1%</td>
<td>20.9%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
Next: Multiple Documents

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