Filling the Blanks (hint: plural noun) for Mad Libs® Humor

Nabil Hossain*, John Krumm**, Lucy Vanderwende**, Eric Horvitz** and Henry Kautz*

*Department of Computer Science
University of Rochester
Rochester, New York

**Microsoft Research AI
Redmond, Washington
Title: The Mona Lisa

Mona Lisa is a/an ________ (adjective) 16th century portrait painted in ________ (type of liquid) by Leonardo da Vinci.

Many people think Mona Lisa’s _____ (noun) is mysterious.
Title: The Mona Lisa

Mona Lisa is a/an __famous___ (adjective) 16th century portrait painted in __oil___ (type of liquid) by Leonardo da Vinci.

Many people think Mona Lisa’s __smile___ (noun) is mysterious.
Mona Lisa is a/an ______ (adjective) 16th century portrait painted in ______ (type of liquid) by Leonardo da Vinci.

Many people think Mona Lisa’s _____ (noun) is mysterious.
Title: The Mona Lisa

Mona Lisa is a/an hilarious (adjective) 16th century portrait painted in lemonade (type of liquid) by Leonardo da Vinci.

Many people think Mona Lisa’s gender (noun) is mysterious.
Title: The Mona Lisa

Mona Lisa is a/an hilarious (adjective) 16th century portrait painted in lemonade (type of liquid) by Leonardo da Vinci.

Many people think Mona Lisa’s gender (noun) is mysterious.

• Incongruity Theory of Humor:
  • Unexpected or surprising word/concept that still fits the sentence
Contributions

- **Fun Libs**: publicly available dataset of Mad Lib-like stories
- Annotation of dataset (using Amazon Mechanical Turk)
  - *qualified* player turkers to fill-in blanks
  - *qualified* judge turkers to grade funniness of story
- Develop **Libitum** algorithm that helps humans generate funnier Mad Libs than they can usually generate on their own
- Analysis of what makes our results humorous
Fun Libs Dataset

• Analyzed 50 original Mad Libs

• Matched these distributions:
  • hint
  • words per hint
  • sentence

<table>
<thead>
<tr>
<th>Hint Type</th>
<th>Mad Libs</th>
<th>Fun Libs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>7.06</td>
<td>7.00</td>
</tr>
<tr>
<td>Adjective</td>
<td>4.06</td>
<td>3.12</td>
</tr>
<tr>
<td>Verb</td>
<td>1.22</td>
<td>3.10</td>
</tr>
<tr>
<td>Part of the Body</td>
<td>0.98</td>
<td>0.46</td>
</tr>
<tr>
<td>Adverb</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>Type of Food</td>
<td>0.22</td>
<td>0.20</td>
</tr>
<tr>
<td>Animal</td>
<td>0.20</td>
<td>0.36</td>
</tr>
<tr>
<td>Type of Liquid</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>All Blanks</td>
<td>16.0*</td>
<td>14.78</td>
</tr>
</tbody>
</table>

• Created 50 Fun Libs from Wikipedia
  • 40 for training
  • 10 for testing
Judges Selection

• Judges shall not consider cheap humor as funny
  • Bathroom humor, slang, sexual references

• Give 7 Mad Lib-like stories to turkers to grade
  • 3 known “not funny” wikipedia articles
  • 4 at least a bit funny

• Grade on the following scale:
  • Not Funny (0), Somewhat Funny (1), Moderately Funny (2), Funny (3)

• Selected 45 judges
Players Selection

- Two sets of experiments:
  - 50 turkers each fill in 2 stories
  - 30 turkers each fill in 2 different stories

- Rank players using qualified judges
  - each story judged by 10 judges
  - Selected 30 players
Filling-in Task

• 50 Mad Libs each filled in by 5 players
Filling-in Task

• 50 Mad Libs each filled in by 5 players

Elephants are the ________ (adjective) living land mammals. An elephant's most obvious part is the ________ (part of the body). An elephant uses its trunk to grab objects such as ________ (type of food). Though the rest of an elephant's hide is ________ (adjective) and thick, its trunk is very ________ (adjective) and ________ (adjective). The elephant usually stands ________ (adverb), raises its trunk, and blows, which is a signal to ________ (animal [plural]) and wildlife. Their ways of acting toward other ________ (animal [plural]) are hard for ________ (animal) to understand. Most elephant sounds are so ________ (adjective) that people cannot hear them, but elephants can hear these sounds far away. Elephants also have ________ (part of the body) coming out of their upper jaws. A lot of ivory comes from elephant tusks. Ivory traders killed many elephants, so ________ (verb ending in "ing") them is now illegal.

Q1: How funny do you think the story is, with the words you entered? Grade on a scale of 0-3 below:

- 0 (Not Funny)
- 1 (Slightly Funny)
- 2 (Moderately Funny)
- 3 (Funny)
Grading Task

• Each filled-in story graded by 9 judges

• Grade individual blanks and overall story
Grading Task

• Each filled-in story graded by 9 judges
• Grade individual blanks and overall story

Title: Synopsis of the Adventures of Tom Sawyer

Tom Sawyer is 12 years old and ☐ casino ☐ (adjective) town along the Mississippi River. Because Tom went to school and made his ☐ drawers ☐ (noun [plural]) dirty in a fight, he gets punished and he has to ☐ paint ☐ (verb) the fence. He does not ☐ want ☐ (verb) to do this, and makes some other ☐ fools ☐ (noun [plural]) do the ☐ boring ☐ (adjective) work. Tom falls in love with Becky Thatcher, the Judge’s ☐ grandmother ☐ (noun). He convinces her to get "engaged" to him but when she finds out that she is not the ☐ first ☐ (adjective) girl he is engaged to, she breaks up with him. One night, Tom goes with Huckleberry Finn to the ☐ clinic ☐ (noun) to try out a cure for ☐ vapors ☐ (noun [plural]). When they ☐ stumble ☐ (verb) there at midnight, they see Injun Joe ☐ tickling ☐ (verb ending in "ing") Dr. Robinson. The murderer ☐ blames ☐ (verb ending in "s") his drunken companion Muff Potter for committing the ☐ crime ☐ (noun) who gets arrested the next day.

Q1: How funny do you think the above Mad Lib is, with the words as entered? Grade on a scale of 0-3:

☐ 0 (Not Funny) ☐ 1 (Slightly Funny) ☐ 2 (Moderately Funny) ☐ 3 (Funny)
Funny Word Classifier

- **funny words**: words receiving 6 or more funny votes
- **not funny words**: words receiving 3 or fewer funny votes
- otherwise word is not used in building classifier

<table>
<thead>
<tr>
<th></th>
<th>Filled-in stories</th>
<th>Filled-in words selected</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td>30 × 5 = 150</td>
<td>1449</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>10 × 5 = 50</td>
<td>490</td>
</tr>
</tbody>
</table>

- Train random forest classifier to predict whether a filled-in word is funny
Features

• Hint type
• Word length
• LM joint probability
• LM joint probability relative position
• Letter trigram probability (min, max, avg)
• Similarity to title, story, sentence (GloVe word vectors)
Funny Word Classifier

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (Funny)</td>
<td>0.712</td>
<td>0.715</td>
</tr>
<tr>
<td>Recall (Funny)</td>
<td>0.856</td>
<td>0.801</td>
</tr>
<tr>
<td>F1 score (Funny)</td>
<td>0.778</td>
<td>0.756</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.727</td>
<td>0.695</td>
</tr>
</tbody>
</table>

Relative Importance

- sentenceSim
- storySim
- titleSim
- letterProbAvg
- letterProbMin
- letterProbMax
- hintType
- wordLength
- jointProbPos
- jointProb
Funny Word Classifier

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (Funny)</td>
<td>0.712</td>
<td>0.715</td>
</tr>
<tr>
<td>Recall (Funny)</td>
<td>0.856</td>
<td>0.801</td>
</tr>
<tr>
<td>F1 score (Funny)</td>
<td>0.778</td>
<td>0.756</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>0.727</strong></td>
<td><strong>0.695</strong></td>
</tr>
</tbody>
</table>

Baseline | Train | Validation |
---------|-------|------------|
Chance   | 0.571 | 0.543      |
Chance Hint | 0.595 | 0.591      |
Linear SVM (3 feat.) | 0.606 | 0.600      |

![Relative Importance Chart](chart.png)
Funnier filled-in words are similar to title but not to story!
Approaches for filling in Mad Libs

- **FreeText:**
  - Players fill in the blanks
Approaches for filling in Mad Libs

- **FreeText:**
  - Players fill in the blanks

- **LMC** (Language Model + Multiple Choice):
  - For each blank:
    - Use a Language Model (LM) to generate candidate words that fit the hint type
    - Keep top 20 words according to LM joint probability
    - Use humans to decide which candidate to fill-in blank with
Approaches for filling in Mad Libs

• **FreeText:**
  • Players fill in the blanks

• **LMC (Language Model + Multiple Choice):**
  • For each blank:
    • Use a Language Model (LM) to generate candidate words that fit the hint type
    • Keep top 20 words according to LM joint probability
    • Use humans to decide which candidate to fill-in blank with

• **Libitum:**
  • For each blank:
    • Use a Language Model (LM) to generate candidate words that fit the hint type
    • **Keep top 20 “funny” words as ranked by funny word classifier**
    • Use humans to decide which candidate to fill-in blank with
Results

- 10 test stories
- for each approach
  - each story filled in by 3 players
  - each filled-in story graded by 9 judges

<table>
<thead>
<tr>
<th>Grade</th>
<th>FreeText</th>
<th>LMC</th>
<th>Libitum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>65</td>
<td>102</td>
<td>37</td>
</tr>
<tr>
<td>1</td>
<td>97</td>
<td>100</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>56</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>44</td>
</tr>
<tr>
<td>mean</td>
<td>1.25</td>
<td>0.92</td>
<td>1.51</td>
</tr>
</tbody>
</table>
Results

- 10 test stories
- for each approach
  - each story filled in by 3 players
  - each filled-in story graded by 9 judges

<table>
<thead>
<tr>
<th>Grade</th>
<th>FreeText</th>
<th>LMC</th>
<th>Libitum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>65</td>
<td>102</td>
<td>37</td>
</tr>
<tr>
<td>1</td>
<td>97</td>
<td>100</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>56</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>44</td>
</tr>
<tr>
<td>mean</td>
<td>1.25</td>
<td>0.92</td>
<td>1.51</td>
</tr>
</tbody>
</table>
Results

- 10 test stories
- for each approach
  - each story filled in by 3 players
  - each filled-in story graded by 9 judges

<table>
<thead>
<tr>
<th>Grade</th>
<th>FreeText</th>
<th>LMC</th>
<th>Libitum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>65</td>
<td>102</td>
<td>37</td>
</tr>
<tr>
<td>1</td>
<td>97</td>
<td>100</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>84</td>
<td>56</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>12</td>
<td>44</td>
</tr>
<tr>
<td>mean</td>
<td>1.25</td>
<td>0.92</td>
<td>1.51</td>
</tr>
</tbody>
</table>

Fully Automated System’s mean funniness grade: 0.80
A beauty contest is a/an ____________ (adj) contest to decide which _________ (n) is the most ____________ (adj)
A beauty contest is a/an \textit{observational} \textbf{(adj)} contest to decide which \textbf{organism} \textbf{(n)} is the most \textit{flammable} \textbf{(adj)}
Title: Beauty Contest

A beauty contest is an observational (adj) contest to decide which organism (n) is the most flammable (adj).

A beauty contest is an elaborate (adj) contest to decide which brawler (n) is the most deadly (adj).
Excerpts

Title: Beauty Contest

Libitum:
A beauty contest is a/an observational (adj) contest to decide which organism (n) is the most flammable (adj)

FreeText:
A beauty contest is a/an elaborate (adj) contest to decide which brawler (n) is the most deadly (adj)

• Human performance better with dependent blanks
Excerpts

Title: Ducks

LMC:
Most ducks are evil (adj) birds, they can be found in both saltwater and fresh coffee (liquid) . . . Some ducks are not evil (adj), and they are bred and kept by jesus (n)

FreeText:
Most ducks are republican (adj) birds, they can be found in both saltwater and fresh vodka (liquid) . . . People commonly feed ducks in ponds stale jokes (n), thinking that the ducks will like to have something to ridicule (verb)
Correlation Analysis

- **Incongruity Theory of Humor:**
  - Unexpected or surprising word/concept that still fits the sentence

- Judges graded how **coherent** each filled-in story was:
  - 0: crazy and disconnected
  - 1: somewhat makes sense
  - 2: hangs together well
  - 3: almost or totally coherent
Correlation Analysis

- **Incongruity Theory of Humor:**
  - Unexpected or surprising word/concept that still fits the sentence

- Judges graded how **coherent** each filled-in story was:
  - 0: crazy and disconnected
  - 1: somewhat makes sense
  - 2: hangs together well
  - 3: almost or totally coherent

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>FText</th>
<th>LMC</th>
<th>Libitum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incongruity</td>
<td>0.638</td>
<td>0.620</td>
<td>0.650</td>
<td>0.515</td>
</tr>
<tr>
<td>Coherence</td>
<td>0.390</td>
<td>0.345</td>
<td>0.023</td>
<td>0.578</td>
</tr>
</tbody>
</table>

*Correlations with funniness grade*
Best “Batman” Story

Batman is a fictional character and one of the most famous / obese / sexiest / inefficient (adjective) superheroes . . . Batman began in comic books and he was later used / liposuctioned / arrested / imprisoned (verb [past]) in several movies, TV programs, and books. Batman lives in the fictional / edible / holy / abyssal (adjective) city of Gotham. . . Batman’s origin story is that as a/an young / obnoxious / adult / algebraic (adjective) child, Bruce Wayne saw a robber murder / eat / play / prank (verb) his parents after the family left a/an theater / sauna / trail / bag (noun). Bruce decided that he did not want that kind of violence / meal / luck / poem (noun) to happen to anyone else. He dedicated his life to protect / devouring / pronounce / demolish (verb) Gotham City. Wayne learned many different ways to fight / nibble / spell / crochet (verb) as he grew up. As an adult, he wore a/an costume / prosthesis / wig / veil (noun) to protect his identity / belly / head / jaw (noun) while fighting crime / gelatin / poverty / acne (noun) in Gotham.

• Key
• Original / FreeText / LMC / Libitum
Conclusion

• Released a new benchmark for computational humor research
  • dataset of filled-in Fun Libs and their funniness grades
• Developed Libitum algorithm that can rank blank-filling words by funniness
• Compared advantages and disadvantages of three approaches for filling in Mad Libs
  • Overall, the computer-aided Libitum approach outperforms human-only approach
  • Some humans generate humor using words that are funny and coherent which is difficult to achieve using Libitum