Evaluation of Text Generation: Automatic Evaluation vs. Variation

Amanda Stent, Mohit Singhai, Matthew Marge
Columbia University
Natural Language Generation

Content Planning

Text Planning

Sentence Planning

Surface Realization
Approaches to Surface Realization

- Template based
  - Domain-specific
  - All output tends to be high quality because highly constrained

- Grammar based
  - Typically one high quality output per input

- Forest based
  - Many outputs per input

- Text-to-text
  - No need for other generation components
Surface Realization Tasks

- To communicate the input meaning as completely, clearly and elegantly as possible by careful:
  - Word selection
  - Word and phrase arrangement
  - Consideration of context
Importance of Lexical Choice

I drove to Rochester.

I went to Rochester.

I raced to Rochester.
Importance of Syntactic Structure

- I picked up my coat three weeks later from the dry cleaners in Smithtown
- In Smithtown I picked up my coat from the dry cleaners three weeks later
Evaluating Text Generators

- Per-generator: Coverage

- Per-sentence:
  - Adequacy
  - Fluency / syntactic accuracy
  - Informativeness

- Additional metrics of interest:
  - Range: Ability to produce valid variants
  - Readability
  - Task-specific metrics
    - E.g. for dialog
Evaluating Text Generators

- Human judgments
- Parsing + interpretation
- Automatic evaluation metrics – for generation or machine translation
  - Simple string accuracy+
  - NIST*
  - BLEU*+
  - F-measure*
  - LSA
What is a “good” sentence?

fluent

adequate

readable
Approach

Question: Which evaluation metric or set of evaluation metrics least punishes variation?
- Word choice variation
- Syntactic structure variation

Procedure: Correlation between human and automatic judgments of variations
- Context not included
Lexical and Syntactic variation

(a) I bought tickets for the show on Tuesday.

(b) It was the show on Tuesday for which I bought tickets.

(c) I got tickets for the show on Tuesday.

(d) I bought tickets for the Tuesday show.

(e) On Tuesday I bought tickets for the show.

(f) For the show on Tuesday tickets I bought.
## String Accuracy

### Simple String Accuracy

### Generation String Accuracy
- \((M + I' + D' + S) / \#Words\) (Rambow et. al. 2002)

<table>
<thead>
<tr>
<th>The</th>
<th>dog</th>
<th>saw</th>
<th></th>
<th>the</th>
<th>man</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>man</td>
<td>was</td>
<td>seen</td>
<td>by</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>S</td>
<td>D</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I, D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I, D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I, D</td>
</tr>
</tbody>
</table>
+ BLEU

- Developed by Papenini et. al. at IBM
- Key idea: count matching subsequences between the reference and candidate sentences
- Avoid counting matches multiple times by clipping
- Punish differences in sentence length

<table>
<thead>
<tr>
<th>The dog saw the man</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man was seen by the dog</td>
</tr>
</tbody>
</table>
NIST ngram

- Designed to fix two problems with BLEU:
  - Geometric mean penalizes large N
  - Might like to prefer ngrams that are more informative = less likely

- Arithmetic average over all ngram co-occurrences

- Weight “less likely” ngrams more

- Use a brevity factor to punish varying sentence lengths
F Measure

- Idea due to Melamed 1995, Turian et. al. 2003

- Same basic idea as ngram measures

- Designed to eliminate “double counting” done by ngram measures

- $F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

- $\text{Precision}(\text{candidate}|\text{reference}) = \frac{\text{maximum matching size(\text{candidate, reference})}}{|\text{candidate}|}$

- $\text{Precision}(\text{candidate}|\text{reference}) = \frac{\text{maximum matching size(\text{candidate, reference})}}{|\text{reference}|}$
<table>
<thead>
<tr>
<th>the</th>
<th>?</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>saw</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>man</td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>

The man was seen by the dog
LSA

- Doesn’t care about word order

- Evaluates how similar two bags of words are with respect to a corpus
  - Measures “similarity” with cooccurrence vectors

- A good way of evaluating word choice?
<table>
<thead>
<tr>
<th>Eval. Metric</th>
<th>Means of measuring fluency</th>
<th>Means of measuring adequacy</th>
<th>Means of measuring readability</th>
<th>Punishes length differences?</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSA</td>
<td>Comparison against reference sentence</td>
<td>Comparison against reference sentence</td>
<td>Comparison against reference sentence from same context*</td>
<td>Yes (punishes deletions, insertions)</td>
</tr>
<tr>
<td>NIST n-gram, BLEU</td>
<td>Comparison against reference sentences -- matching n-grams</td>
<td>Comparison against reference sentences</td>
<td>Comparison against reference sentences from same context*</td>
<td>Yes (weights)</td>
</tr>
<tr>
<td>F measure</td>
<td>Comparison against reference sentences -- longest matching substrings</td>
<td>Comparison against reference sentences</td>
<td>Comparison against reference sentences from same context*</td>
<td>Yes (length factor)</td>
</tr>
<tr>
<td>LSA</td>
<td>None</td>
<td>Comparison against word co-occurrence frequencies learned from corpus</td>
<td>None</td>
<td>Not explicitly</td>
</tr>
</tbody>
</table>
Experiment 1

Sentence data from Barzilay and Lee’s paraphrase generation system (Barzilay and Lee 2002)

- Includes word choice variation, e.g.
  - Another person was also seriously wounded in the attack vs
  - Another individual was also seriously wounded in the attack

- Includes word order variation, e.g.
  - A suicide bomber blew himself up at a bus stop east of Tel Aviv on Thursday, killing himself and wounding five bystanders, one of them seriously, police and paramedics said
  - A suicide bomber killed himself and wounded five, when he blew himself up at a bus stop east of Tel Aviv on Thursday
Paraphrase Generation

1. Cluster like sentences
   - By hand or using word n-gram co-occurrence statistics
   - May first remove certain details

2. Compute multiple-sequence alignment
   - Choice points and regularities in input sentence pairs/sets in a corpus

3. Match lattices
   - Match between corpora

4. Generate
   - Lattice alignment
Paraphrase Generation Issues

- Sometimes words chosen for substitution carry unwanted connotations.
- Sometimes extra words are chosen for inclusion (or words removed) that change the meaning.
Discussion

- These metrics achieve some level of correlation with human judgments of adequacy
  - But could we do better?

- Most metrics are negatively correlated with fluency
  - Word order or constituent order variation requires a different metric

- Automatic evaluation metrics other than LSA punish word choice variation

- Automatic evaluation metrics other than LSA punish word order variation
  - Are not as effected by word order variation as by word choice variation
  - Punish legitimate and illegitimate word and constituent reorderings equally
Discussion

- **Fluency**
  - These metrics are not adequate for evaluating fluency in the presence of variation

- **Adequacy**
  - These metrics are barely adequate for evaluating adequacy in the presence of variation

- **Readability**
  - These metrics do not claim to evaluate readability
A Preliminary Proposal

- Modify automatic evaluation metrics as follows:
  - Not punish legitimate word choice variation
    - E.g. using WordNet
    - But the ‘simple’ approach doesn’t work
  - Not punish legitimate word order variation
    - But need a notion of constituency
When using metrics that depend on a reference sentence, use
- A set of reference sentences
  - Try to get as many of the word choice and word order variations as possible in the reference sentences
- Reference sentences from the same context as the candidate sentence
  - To approach an evaluation of readability
- And combine with some other metric for fluency
  - For example, a grammar checker
To evaluate a generator:
- Evaluate for coverage using recall or related metric
- Evaluate for ‘precision’ using separate metrics for fluency, adequacy and readability
  - At this point in time, only fluency may be evaluable automatically, using a grammar checker
  - Adequacy can be approached using LSA or related metric
  - Readability can only be evaluated using human judgments at this time
Current and Future Work

- Other potential evaluation metrics:
  - F measure plus WordNet
  - Parsing as measure of fluency
  - F measure plus LSA
  - Use multiple-sequence alignment as an evaluation metric

- Metrics that evaluate readability