Evaluation Metrics for Natural Language Generation

Srinivas Bangalore

srini@research.att.com

Joint work with Owen Rambow and Steve Whittaker

AT&T Labs – Research

Statistical Generation Day,
October 9, 2001

Evaluation in NLP

- Evaluation of systems and technologies have had significant impact on NLP in the past decade.
- Evaluation metrics
  - help identify direction for development effort
  - help in cross-system comparisons
  - But: Research efforts may be limited by evaluation metrics.
- Applies to trainable and hand-crafted technologies.

Evaluation in NLP

- Evaluation of systems under the HLT, MUC, TREC, TIDES programs
- Evaluation of technologies
  - SpeechEval (ATIS, Switchboard, Broadcast News)
  - POSEval (unofficial in US, more official in Europe)
  - ParseEval
  - SenseEval
- Evaluation metrics are hard to come by for output technologies.
  - Machine Translation
  - NL Generation
  - Speech Synthesis
  - Dialog Systems
- Why?

Evaluation Metrics for NL Generation

- Trainable Generation system: Training and Test loop
- Metric needed for developing stochastic generator:
  - objective and automatic
  - without human intervention
  - quick turnaround
- These metrics were not intended to compare realizers (but . . .)
- In the context of surface realizer, accuracy is measured against a reference string.

Two String-Based Evaluation Metrics

- String edit distance between reference string and result string (length in words: R)
  - Substitutions (S)
  - Insertions (I)
  - Deletions (D)
  - Moves = pairs of Deletions and Insertions (M)
  - Remaining Insertions (I') and Deletions (D')

- Example:

There	was	no	cost	estimate	for		the	second		phase
There	was			estimate	for	phase	the	second	no	cost
.	.	d	d	.	.	i	.	.	i	s

- **Simple String Accuracy** $= (1 - \frac{I+D+S}{R})$

- **Generation String Accuracy** $= (1 - \frac{M+I'+D'+S}{R})$

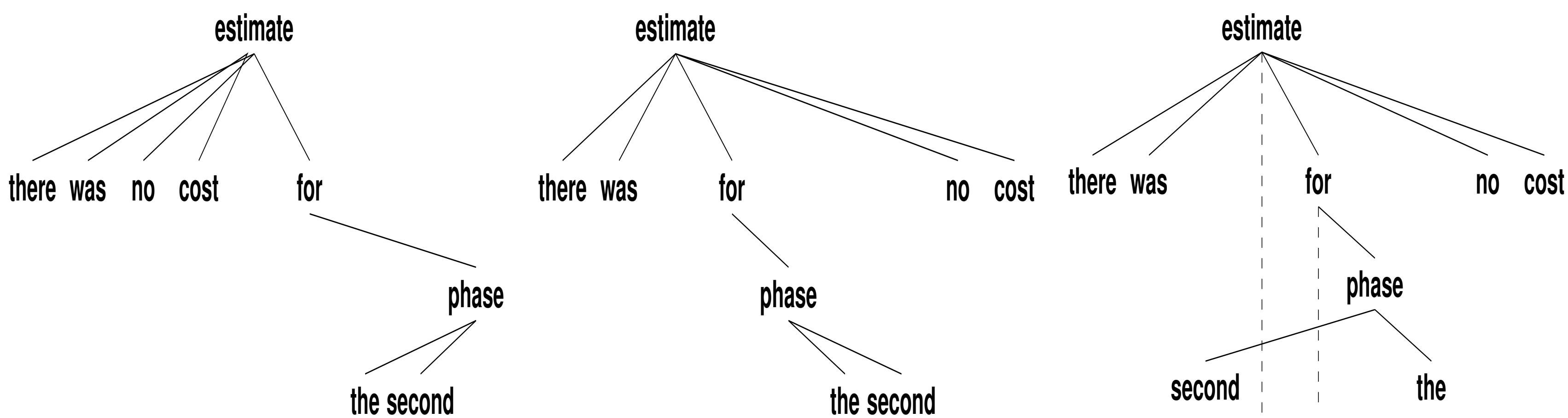
Experiments and Evaluation

- Training corpus: One million words of WSJ corpus
- Test corpus:
  - 100 randomly chosen sentences
  - average sentence length 16.7 words

Model	Generation Accuracy
Baseline	0.562
TM-LM	0.668
TM-XTAG	0.684
TM-XTAG-LM	0.724

Two Tree-Based Evaluation Metrics

- Not all moves equally bad: moves which permute nodes in tree better than moves which "scramble" tree (projectivity)
- **Simple Tree Accuracy** metrics: calculate S, D, I on each treelet
- **Generation Tree Accuracy** metrics: calculate S, M, D', I' on each treelet
- Example: There was estimate for phase the second no cost

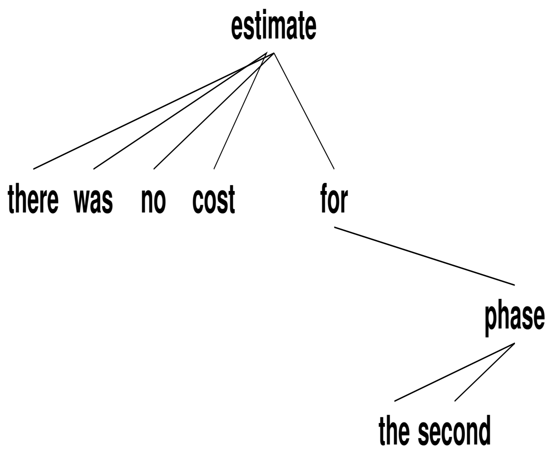


there was no cost estimate for the second phase

there was estimate for phase the second no cost

there was second estimate for phase the no cost

Details of Tree Metric Computation



the second phase
for phase
there was no cost estimate for

Result: there was estimate for **phase the second** no cost

there was estimate **for phase** no cost

there was estimate for no cost

Errors = Insertions=3 + Deletions=3 (Moves=3)
Metric does not need a tree representation for the generated sentence.

• Comparing the Evaluation Metrics

- Example (repeated):

There	was	no	cost	estimate	for		the	second		phase
There	was			estimate	for	phase	the	second	no	cost
.	.	d	d	.	.	i	.	.	i	s

Metric	Simple String Acc	Generation String Acc	Simple Tree Acc	Generation Tree Acc
Tot. # of tokens	9	9	9	9
Unchanged	6	6	6	6
Substitutions	1	1	0	0
Insertions	2	1	3	0
Deletions	2	1	3	0
Moves	0	1	0	3
Tot. # of S, I, D, M	5	4	6	3
Score	0.44	0.56	0.33	0.67

Measuring Performance Using Evaluation Metrics

- Baseline: randomly assigned dependency structure, learn position of dependent to head
- Training corpus: One million words of WSJ corpus
- Test corpus:
  - 100 randomly chosen sentences
  - average sentence length 16.7 words

Tree Model	Simple String Acc	Generation String Acc	Simple Tree Acc	Generation Tree Acc
Baseline LR Model	0.41	0.56	0.41	0.63
FERGUS	0.58	0.72	0.65	0.76

Experimental Validation

- Problem: how are these metrics *motivated*?
- Solution (following Walker et al 1997):
  - Perform experiments to elicit human judgments on sentences
  - Relate human judgments to metrics

Experimental Setup

- Web-based
- Human subjects read short paragraph from WSJ and three or five variants of last sentence constructed by hand
- Humans judge:
  - **Understandability**: How easy is this sentence to understand?
  - **Quality**: How well-written is this sentence?
- Values: 1-7; 3 values have qualitative labels
- Ten subjects; each subject made a total of 24 judgments
- Data normalized by subtracting mean for each subject and dividing by standard deviation; then each variant averaged over subjects

Results of Experimental Validation

- Strong correlations between normalized understanding and quality judgments ($r_{(22)} = 0.94$, $p < 0.0001$)
- The two tree-based metrics correlate with both understandability and quality.
- The string-based metrics do not correlate with either understandability or quality.

Corr. with	Simple String Acc	Generation String Acc	Simple Tree Acc	Generation Tree Acc
Norm. und.	0.08	0.23	**0.51**	**0.48**
Norm. qual.	0.16	0.33	**0.45**	**0.42**

Experimental Validation: Finding Linear Models

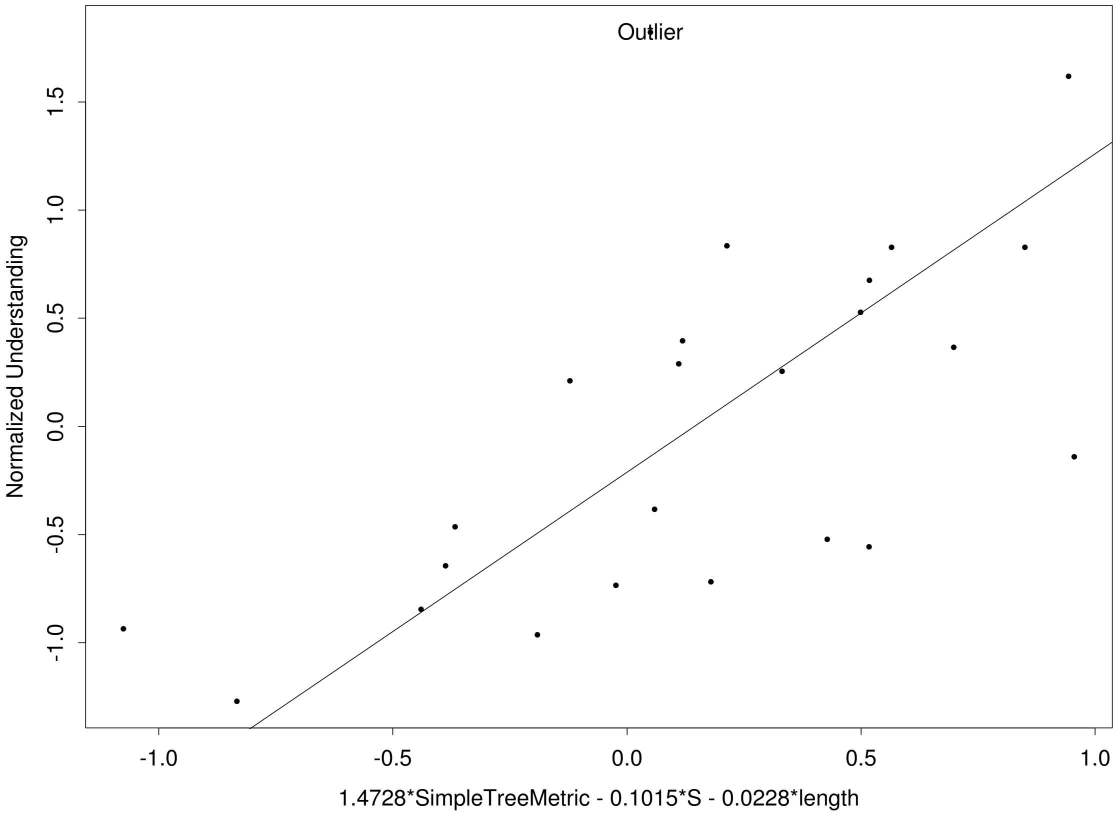
- Other goal of experiment: find better metrics
- Series of linear regressions
  - Dependent measures: normalized understanding and quality
  - Independent measures: different combinations of:
    - The four metrics
    - Sentence length
    - The "problem" variables (S, I, D, M, I', D')
- One outlier excluded from data set
- Can improve on explanatory power of original four metrics

Experimental Validation: Linear Models

Model	User Metric	Exp. Pwr. (R^2)	Stat. Sig. (p value)
Simple String Acc.	Und.	0.02	0.571
Generation String Acc.	Und.	0.02	0.584
Simple Tree Acc.	Und.	0.36	0.003
Generation Tree Acc.	Und.	0.35	0.003
Simple Tree Acc. + S	Und.	0.48	0.001
Simple Tree Acc. + S	Qual.	0.47	0.002
Simple Tree Acc. + M	Und.	0.38	0.008
Simple Tree Acc. + M	Qual.	0.34	0.015
Simple Tree Acc. + Length	Und.	0.40	0.006
Simple Tree Acc. + Length	Qual.	0.42	0.006
Simple Tree Acc. + S + Length	Und.	0.51	0.003
Simple Tree Acc. + S + Length	Qual.	0.53	0.002

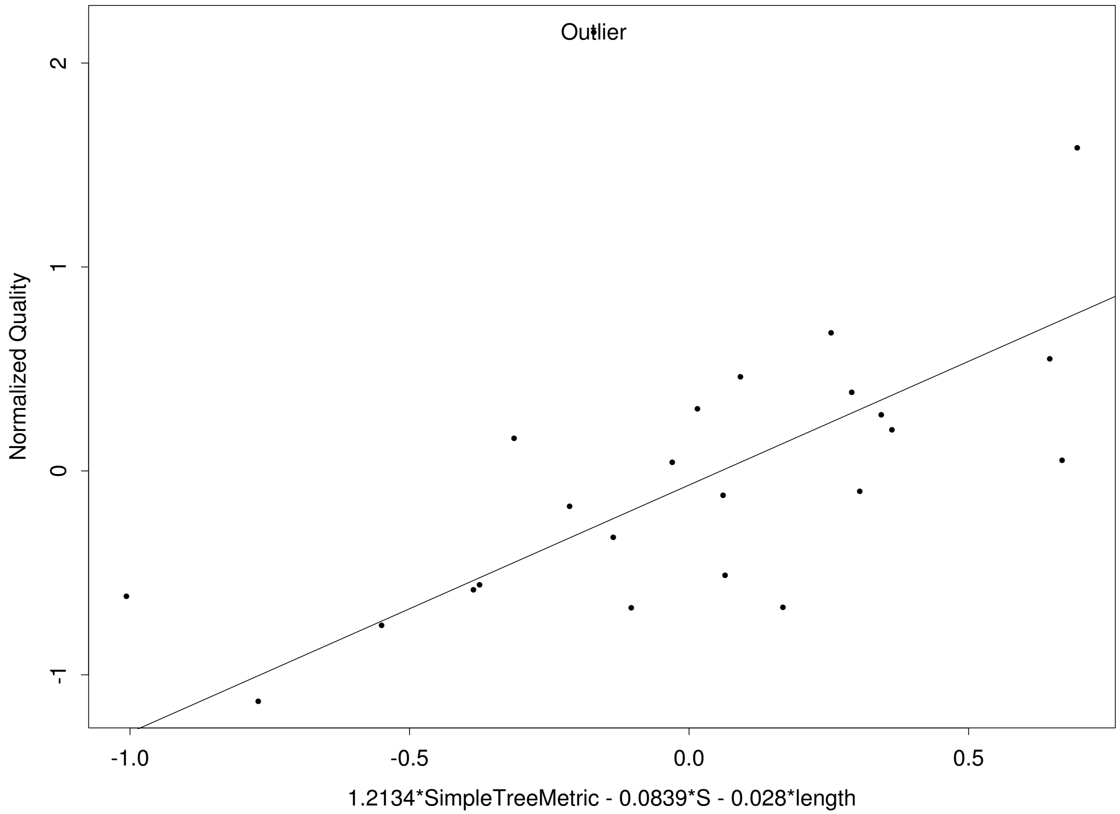
Experimental Validation: Model of Understanding

- Normalized understanding = 1.4728*simple tree accuracy - 0.1015*substitutions - 0.0228 * length - 0.2127.



Experimental Validation: Model of Quality

- Normalized quality = 1.2134*simple tree accuracy - 0.0839*substitutions - 0.0280 * length - 0.0689.



Two New Metrics

- Don't want length to be included in metrics
- **Understandability Accuracy** = (1.3147*simple tree accuracy
  -0.1039*substitutions
  -0.4458) / 0.8689
- **Quality Accuracy** = (1.0192*simple tree accuracy
  -0.0869*substitutions
  -0.3553) / 0.6639
- Scores using new metrics:

Tree Model	Understandability Accuracy	Quality Accuracy
Baseline	-0.08	-0.12
Supertag-based	0.44	0.42

Problems with Evaluation Metrics

The Question of "Gold Standard" in Generation

- There are many ways of saying something
- But: given contextual and genre restrictions, often not that many ways
- Nonetheless, comparison to single reference sentence problematic
- Justification: in stochastic generation, we learn from a corpus because we want to mimic it as closely as possible
- Issues:
  - We are only evaluating word order; word order variation in English is different from other languages
  - We are not taking context into account

Summary

- Need immediate evaluation of performance for development
- Two new metrics which are validated experimentally
- Can also use to compare two different surface realizers
- Ultimate evaluation of a realizer is (probably) in a task-based evaluation of a larger system.

Discussion Topics: Evaluation Metrics

- What about a corpus of paraphrases?
  - Notion of paraphrase: Functional (dialog act), lexico-syntactic, ...
  - Not necessarily naturally occuring, more like a test suite (TSNLP for parsing)
  - Relates to internal and cross-system evaluation
  - Metric for comparing paraphrases (= evaluation metric for NLG)

Discussion Topics: Evaluation Metrics

- Why do we evaluate?
- What do we evaluate?
  - things that are annotated in a corpus
  - user experience
- How do we evaluate?
  - Component vs end-to-end
  - Glass-box vs Black-box
- Relevance of human judgements to metrics
- Relevance of metrics to human judgements
- What should human judgements be about?

Discussion Topics: NLG Issues in Applications

- How to choose among NLG approaches: Rule-based vs Template vs Stochastic NLG.
- Possible metric to choose an approach: Perplexity?
- Rapid prototyping: corpus-based NLG might win.
- End-to-End evaluation.
- How much of stochastic parsing has made it into applications, anyway?

Discussion Topics: Corpus Annotation

Separation of corpus annotation from how the corpus is used

- What phenomena are suitable for corpus-based analysis?

  Higher NLG tasks (Sentence Planning and Text Planning) more difficult to encode.

  - Higher NLG tend to be more application-specific and hence arriving at a annotation standard is difficult.
  - Issues of consensus, annotation standard, knowledge about phenomena: guidelines for inter-annotator agreement require deep understanding of issue.
  - On-going work on dialog annotation and discourse annotation (Marcu).

- What kinds of annotations are needed?
- Can we reuse corpora created for training parsers and word-sense disambiguation models?