Investigating Why BLEU Penalizes Non-Statistical Systems

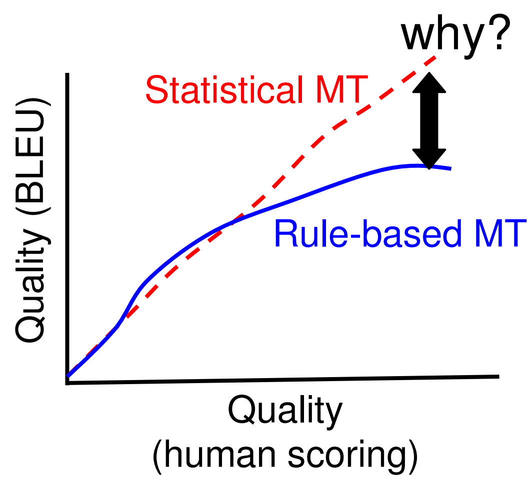
Eduard Hovy
USC Information Sciences Institute
hovy@isi.edu

USC
INFORMATION
SCIENCES
INSTITUTE

Observation

- In DARPA's GALE program, Program manager Joe Olive is worried by this fact:

  BLEU (and similar automated evaluation systems) have a tendency to penalize non-statistical MT engines unfairly as the quality goes up:
  - for better translation, the BLEU score for statistical systems more or less correlates with humans' intuitive judgments,
  - but the BLEU score on rule-based MT systems is artificially low



Possible reasons

1. Degree of divergence from input word order:
   - ngram-based systems follow the input text word sequence rather slavishly; rule-based systems do not
   - The rearrangements (relativization, passiviation, clause reordering, etc.) may not be wrong, but may not be what the gold standard contains
   - So, the more flexible systems are penalized by BLEU
2. Generality of output formulation:
   - Human rule-writers create rules that produce somewhat general outputs to cover multiple closely-related input variations (this reduces their work), while statistical systems learn every little variation separately, in its own peculiarities
   - BLEU scores lower the more general (but not incorrect) translation against the gold-standard texts that are probably more specific, BLEU scores higher the statistical systems' outputs, which are more specific
   - But the rule-based output reads fine, and in some cases better even than the statistical output

USC INFORMATION SCIENCES INSTITUTE

Action

- Joe convened a meeting in May 2007
  - Liz Boschee (BBN), Marjorie Freedman (BBN), Eduard Hovy (ISI), Kevin Knight (ISI), Daniel Marcu (ISI), Mitch Marcus (UPenn), Ralph Weischedel (BBN)
- Question: Can we somehow use more-flexible (syntactic, even semantic) information to recognize correctness of less literal translations?
  - How to encode 'equivalent' syntactic transformations?
  - How to obtain semantic version of input?
  - What are 'equivalent' semantic transformations?

Decision

- BBN will use its Distillation engine to score system outputs against gold standard fragments
- Distillation engine:
  - Runs after IR has located potentially relevant text passages to answer input question
  - Purpose: identify redundancies and irrelevant fragments and produce ranked list of most-relevant fragments
- Distillation engine operation:
  - Produces parse trees and/or fragments
  - Compares them, accepting certain tree transformations
  - Includes some simple paraphrase matching

All work done by Liz Boschee, BBN

USC INFORMATION SCIENCES INSTITUTE

Experiment

- Data: GALE 2006 AGILE HTER texts (4762 sentences)
  - For each document: hypothesis (system output from the AGILE MT system), reference translation, HTER-reference translation (the translation generated during the HTER scoring process)
  - For each sentence: TER and HTER scores

Translation error rate
Human transl error rate

- Matches:
  - **full match**: how well tree A instantiates into tree B
  - **subtree match**: how well the subtrees of tree A instantiate into tree B
  - **node match**: how well the nodes of tree A instantiate into tree B
- For each pair of reference and hypothesis sentences, 6 scores:
  - full match: $hyp \rightarrow ref$ and $ref \rightarrow hyp$
  - subtree match: $hyp \rightarrow ref$ and $ref \rightarrow hyp$
  - node match: $hyp \rightarrow ref$ and $ref \rightarrow hyp$
- Scoring: 4 averages:
  - full match / subtree match / node match / all match average

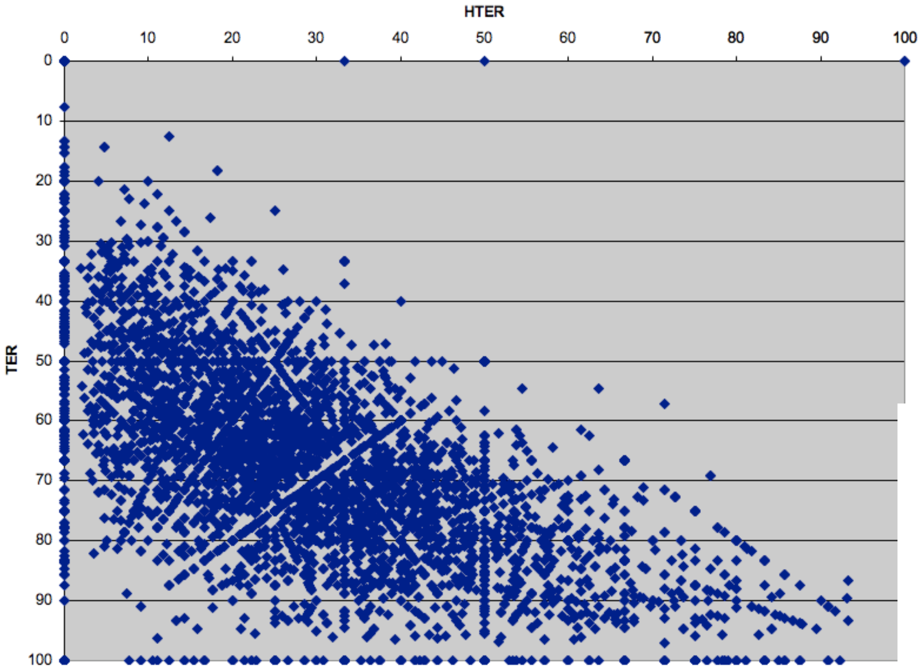
USC INFORMATION SCIENCES INSTITUTE

Checking validity

- Pearson's r score for the correlation of each measure with the HTER scores:
  - full match average: -0.29
  - subtree match average: -0.47
  - node match average: -0.54
  - all match average: -0.50
  - (TER: 0.53)
  - (TER + parser proptrees: -.061)

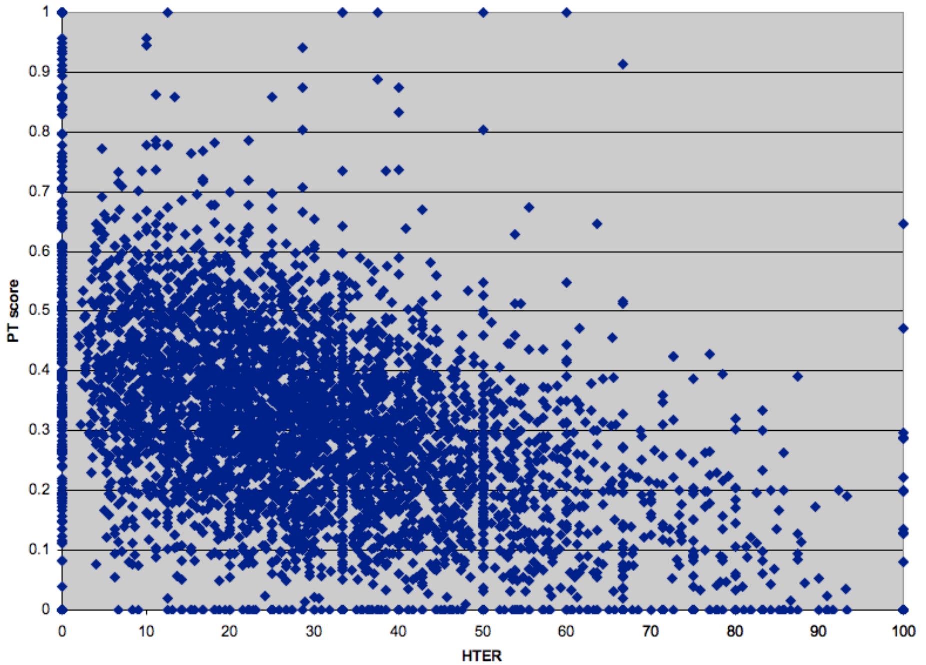
USC INFORMATION SCIENCES INSTITUTE

Findings 1

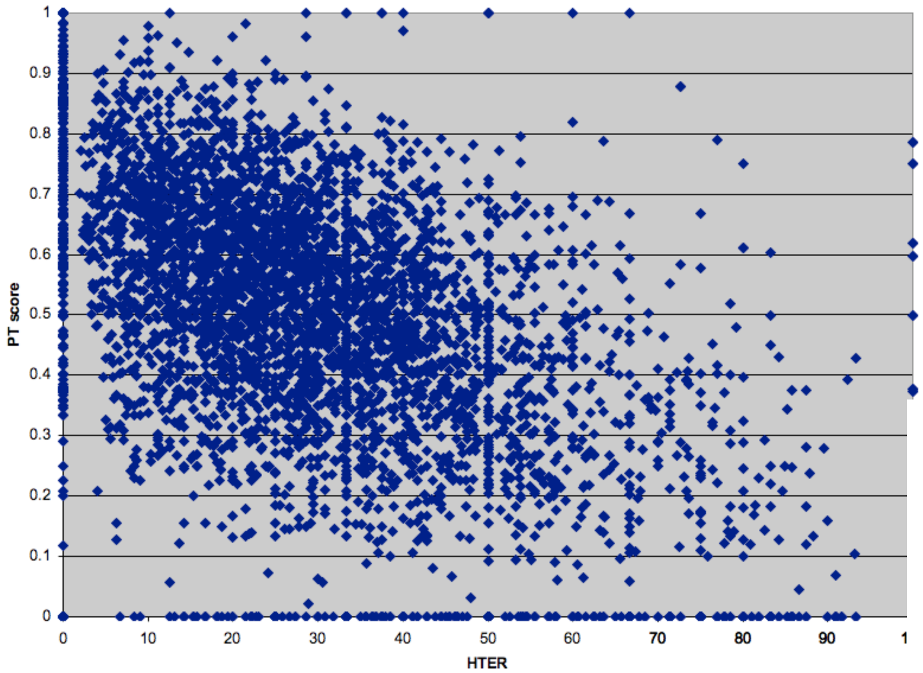


HTER vs TER

HTER vs average

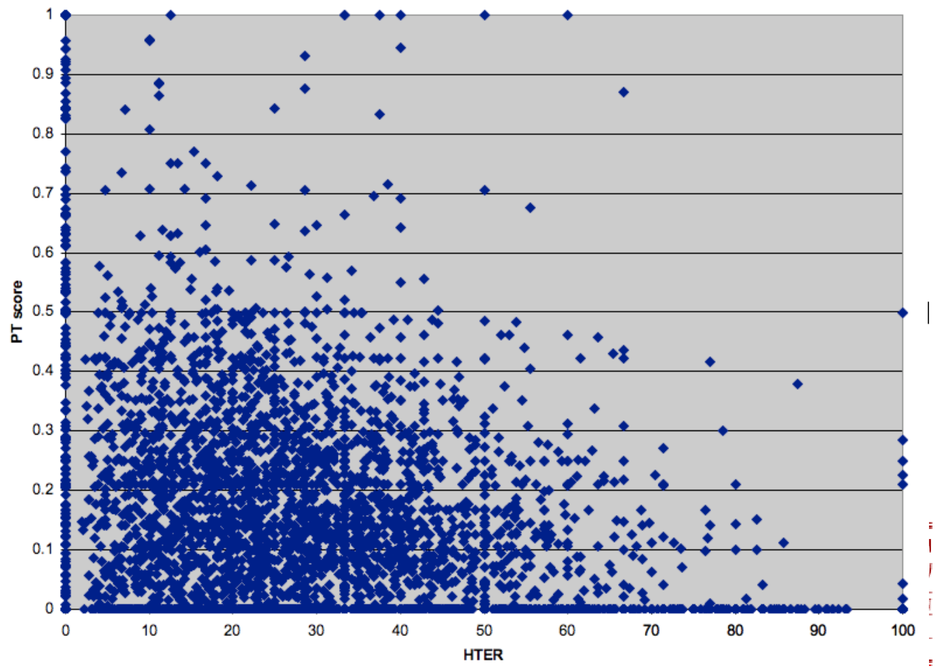


Findings 2



HTER vs nodes

HTER vs full tree



Next steps

- Verify statistics of significance, etc.
- What do the results show? — Draw conclusions and implications
- Define additional eval system parallel to BLEU (?)