
Re-evaluating the Role of BLEU in Machine Translation Research

Chris Callison-Burch,
Miles Osborne and Philipp Koehn

April 7, 2006



Talk Overview

- How do we currently evaluate MT research?
- What assumptions does our methodology rely on?
- Are those assumptions valid?
- If not, what does that imply for the field?

Conducting Research in MT

- Posit theory of how to improve translation quality
- Change the behavior of a translation system accordingly
- Translate a set of test sentences
- Compare translations before and after change
- If **better**, then write a paper

Determining Goodness

- To determine if translation improved, we need to measure translation quality
- Can be done manually by judging a translation's **fluency** and **adequacy**

Fluency

5. Flawless English
4. Good English
3. Broken English
2. Disfluent
1. Incomprehensible

Adequacy

5. All
4. Most
3. Much
2. Some
1. None

Human v. Automatic Evaluation

- Human evaluation is accurate, but
 - It's time consuming
 - It's expensive
 - It's not easy to re-use
- We would like an automatic metric
 - Which can be run quickly at no cost
 - Which correlates with human judgments
- Accomplished by comparing to references

Difficulties of Automatic Evaluation of MT

- Different than Word Error Rate metric used in speech recognition
 - WER assumes a single authoritative reference
 - WER assumes linear ordering
- By contrast, translation has a range of possible realizations
 - A variety of equally valid wordings
 - Some phrases can be moved

Enter: BLEU

- “Bi-Lingual Evaluation Understudy”
- Allows multiple reference translations as an attempt to model the variety of possible translations
- Matches n-grams from reference without putting explicit constraints on order
- Has been shown to **correlate with human judgments** of translation quality

BLEU Detailed

References:

Rodriguez seemed quite calm as he was being led to the American plane that would take him to Miami in Florida .

Rodriguez appeared calm as he was being led to the American plane that was to carry him to Miami in Florida .

Rodriguez appeared calm as he was led to the American plane which will take him to Miami , Florida .

Rodriguez appeared calm while being escorted to the plane that would take him to Miami , Florida .

Hypothesis:

Appeared calm when he was taken to the American plane , which will to Miami , Florida .

Matches:

1-grams:

2-grams:

3-grams:

4-grams:

BLEU Detailed

References:

Rodriguez seemed quite calm as he was being led to the American plane that would take him to Miami in Florida .

Rodriguez appeared calm as he was being led to the American plane that was to carry him to Miami in Florida .

Rodriguez appeared calm as he was led to the American plane which will take him to Miami , Florida .

Rodriguez appeared calm while being escorted to the plane that would take him to Miami , Florida .

Matches:

1-grams: **15**

2-grams:

3-grams:

4-grams:

Hypothesis:

Appeared calm when he was taken to the American plane , which will to Miami , Florida .

BLEU Detailed

References:

Rodriguez seemed quite calm as he was being led to the American plane that would take him to Miami in Florida .

Rodriguez appeared calm as he was being led to the American plane that was to carry him to Miami in Florida .

Rodriguez appeared calm as he was led to the American plane which will take him to Miami , Florida .

Rodriguez appeared calm while being escorted to the plane that would take him to Miami , Florida .

Matches:

1-grams: **15**

2-grams: **10**

3-grams:

4-grams:

Hypothesis:

Appeared calm when he was taken to the American plane , which will to Miami , Florida .

BLEU Detailed

References:

Rodriguez seemed quite calm as he was being led [to the American plane](#) that would take him to Miami in Florida .

Rodriguez appeared calm as he was being led to the American plane that was to carry him to Miami in Florida .

Rodriguez appeared calm as he was led to the American plane which will take him [to Miami , Florida](#) .

Rodriguez appeared calm while being escorted to the plane that would take him to Miami , Florida .

Matches:

1-grams: **15**

2-grams: **10**

3-grams: **7**

4-grams: **3**

Hypothesis:

Appeared calm when he was taken [to the American plane](#) , which will [to Miami , Florida](#) .

BLEU Detailed

- Calculates **n-gram precision** p_n for $n = 1, 2, 3, 4 \dots$ by summing over n-gram matches for every hypothesis translation in test set
- Uses **brevity penalty** to compensate for lack of recall by penalizing translations that are too short

$$\text{BP} = \begin{cases} 1 & \text{if } h > r \\ e^{1-r/h} & \text{if } h \leq r \end{cases}$$

- Bleu is defined as weighted geometric average of p_n offset by BP

$$\text{Bleu} = \text{BP} * \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

Common Assumptions About BLEU

- Bleu is commonly reported as **sole evidence** of improved translation quality in conference papers
- Sometimes failure to improve Bleu is taken as failure to improve translation quality (see “Word Sense Disambiguation v. SMT”)
- This relies on two key assumptions:
 - Accurately accounts for allowable variation in translation
 - Correlates with human judgments

Are These Assumptions Valid?

- Does an improvement in Bleu score guarantee a genuine translation improvement?
- Does a failure to improve Bleu always mean that translation quality has not improved?

Not Always

- We show that in some cases a higher Bleu score is **neither sufficient nor necessary** to ensure genuine translation improvement.
- We do this in two ways
 1. By showing that Bleu has poor model of allowable variation and fails to distinguish between translations of differing quality
 2. By showing two significant counterexamples to Bleu's correlation with human judgments

Equally Scoring Translations: Permutations

- Because Bleu does not constrain order of n-grams, we can construct equal scoring translations by permuting around **bigram mismatch** points

Appeared calm | when | he was | taken | to the American plane | , | which will | to Miami , Florida .

- So this and **40,320** other candidates receive the same score:

which will | he was | , | when | taken | Appeared calm | to the American plane | to Miami , Florida .

- Current systems produce translations with millions of similarly scoring permutations, up to 10^{73} . Likely to be judged equally valid?

Equally Scoring Translations: Substitutions

- Different items may be drawn from references and receive the same score

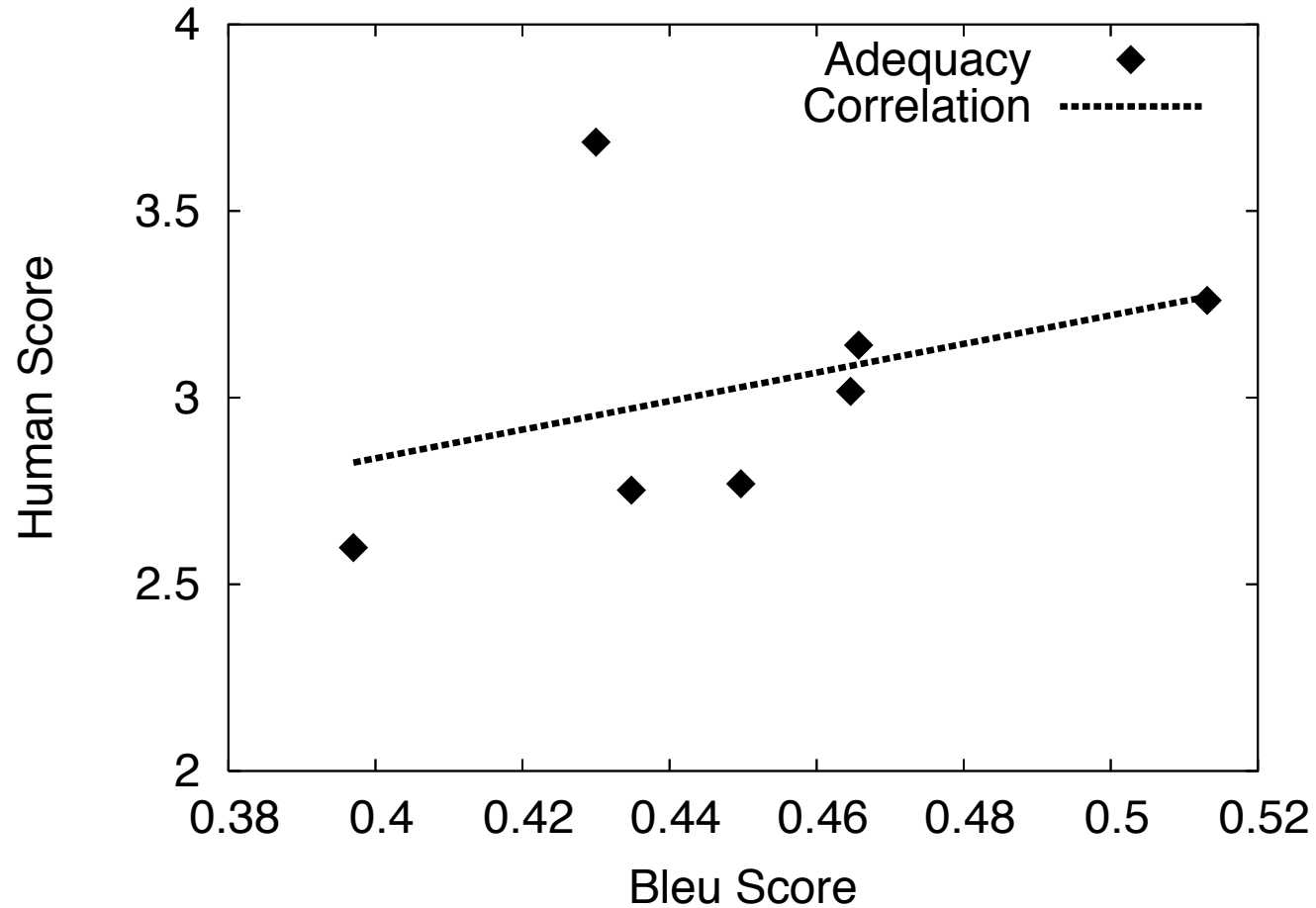
was being led to the | calm as he was | would take | carry him | seemed
quite | when | taken

- Unmatched words (when, taken) can be replaced by anything (black, helicopters)
- Bleu's model of allowable variation in translation is **insufficient** to distinguish between different quality translations
- Bleu cannot be guaranteed to correlate with human judgments

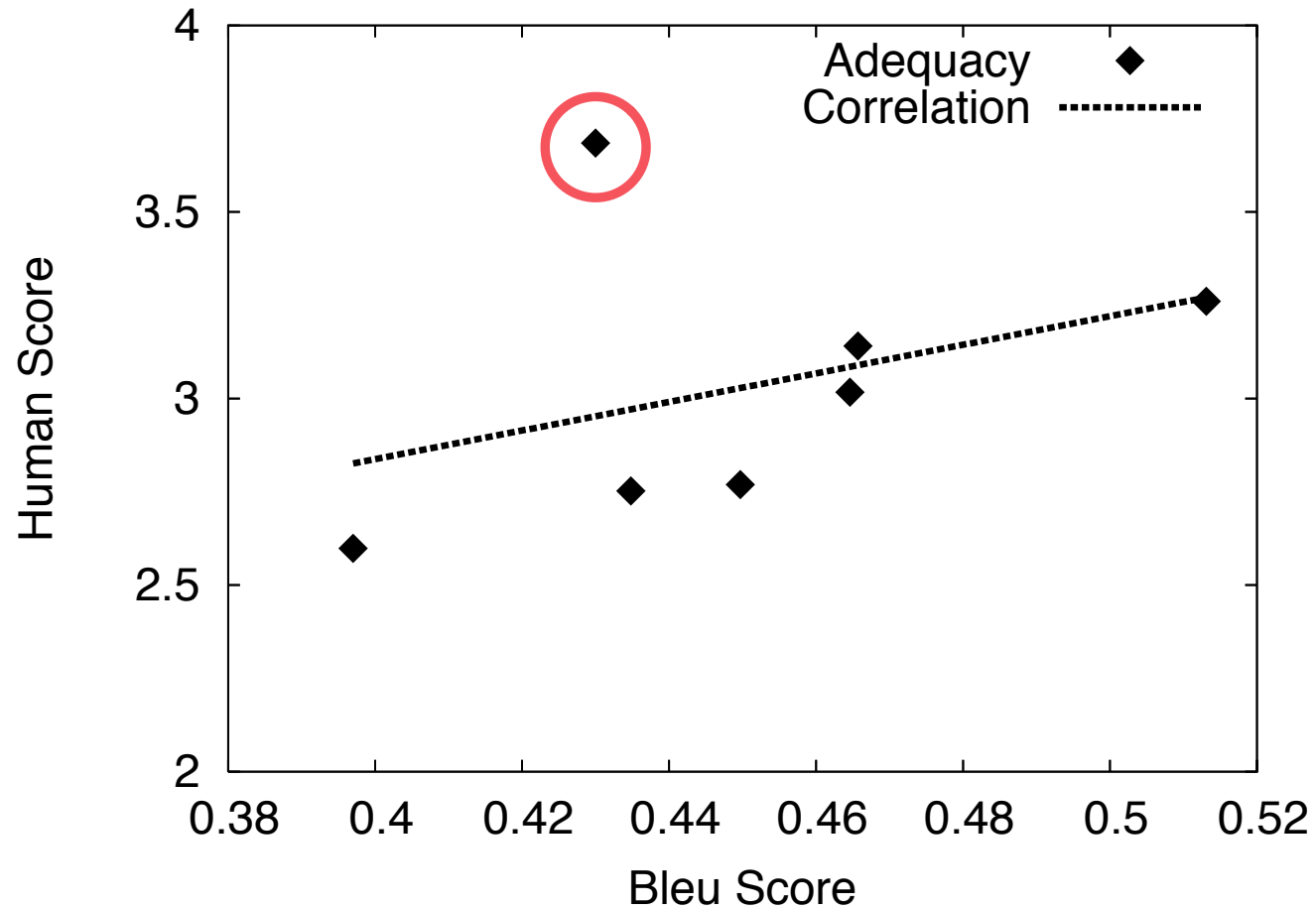
Failures in Practice

- Criticism: Those are *constructed* examples, Bleu assumes cooperative environment
- These failures happen in practice too
 - In the 2005 NIST MT Eval, the *6th* ranked Bleu system scored *1st* in the manual human evaluation
 - Bleu incorrectly ranks poor phrase-based MT system higher than good rule-based system

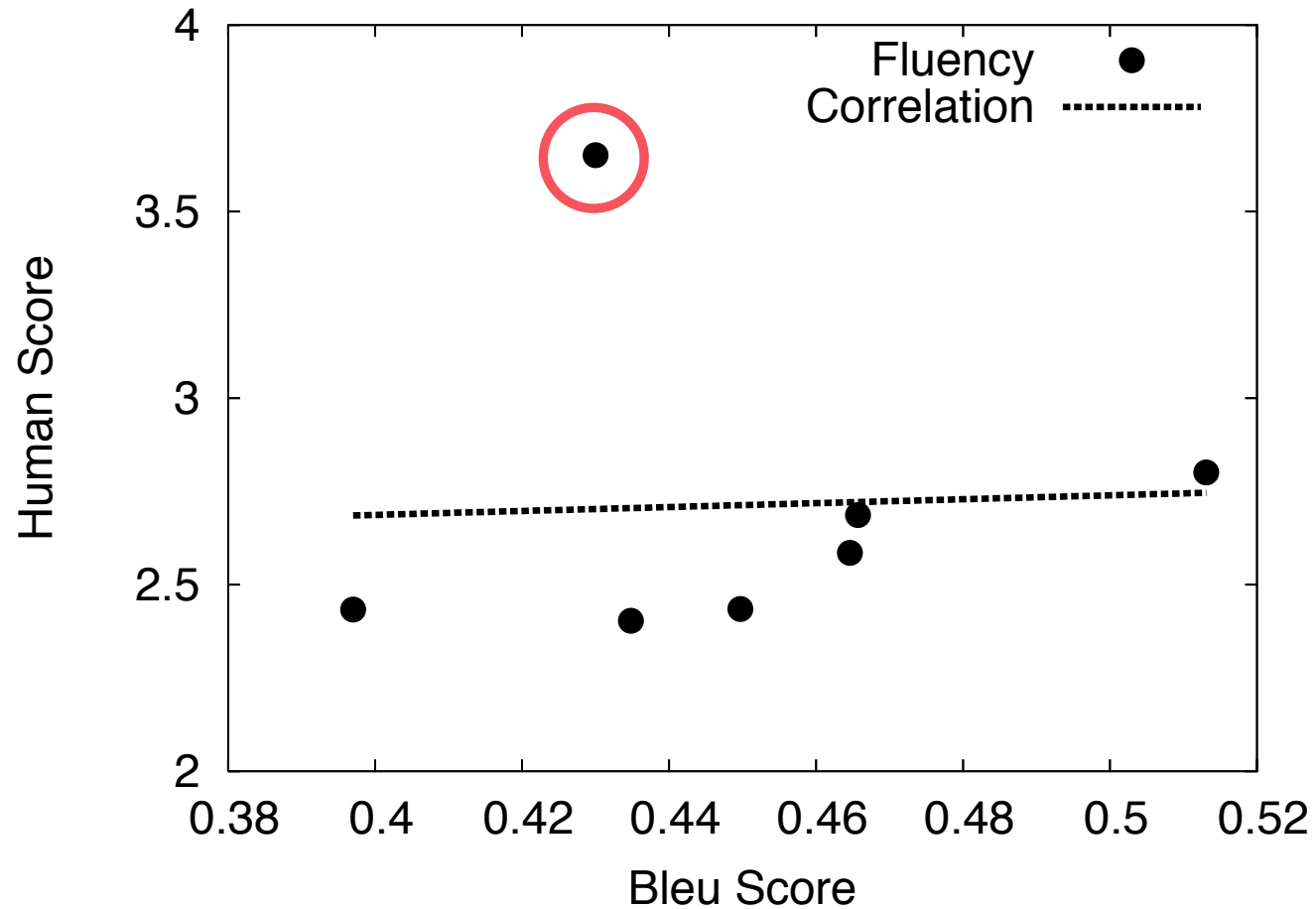
NIST 2005 Results



NIST 2005 Results



NIST 2005 Results



Example

Reference: Iran had already announced Kharazi would boycott the conference after Jordan's King Abdullah II accused Iran of meddling in Iraq's affairs.

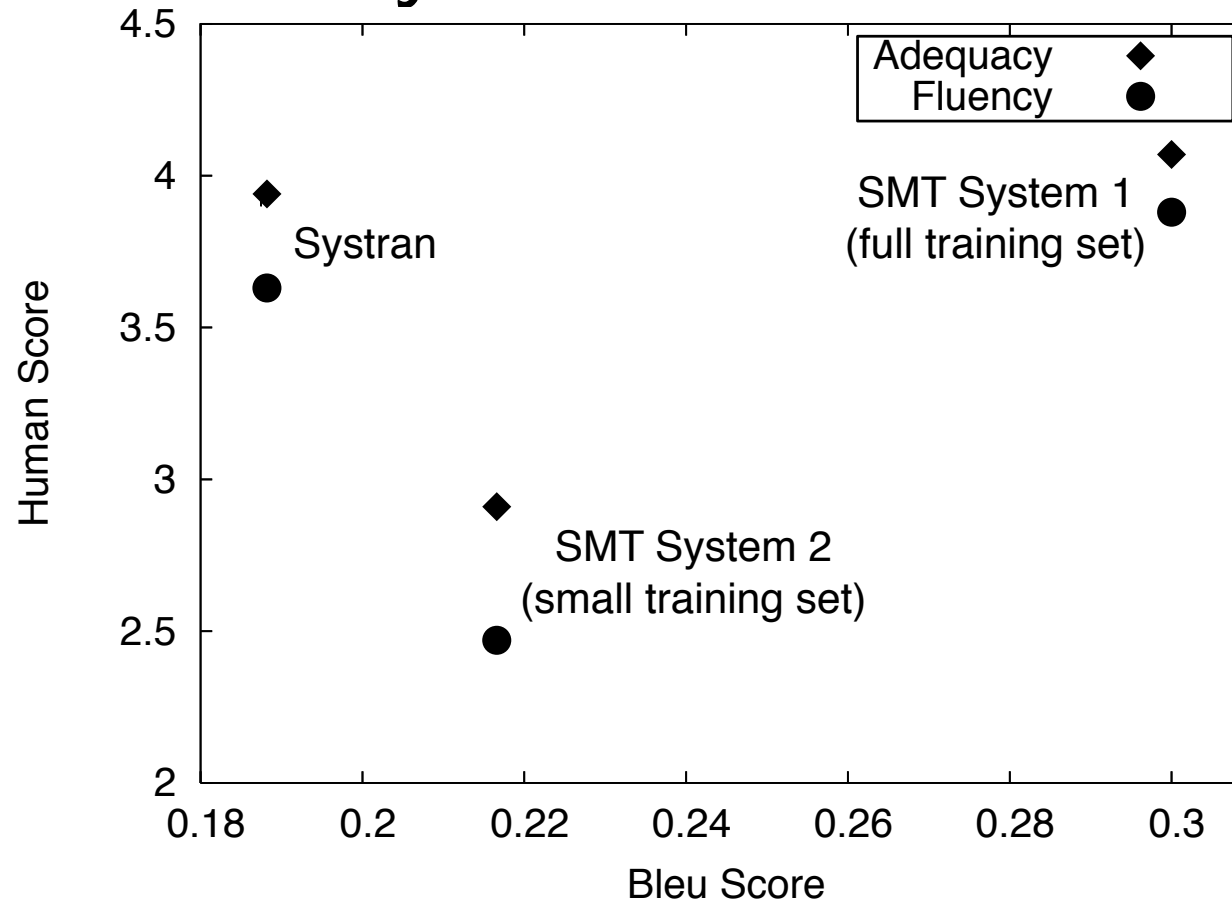
Hypothesis 1: Iran has already stated that Kharazi's statements to the conference because of the Jordanian King Abdullah II in which he stood accused Iran of interfering in Iraqi affairs.

n-gram matches: 27 unigrams, 20 bigrams, 15 trigrams, and ten 4-grams
human scores: Adequacy:3,2 Fluency:3,2

Hypothesis 2: Iran already announced that Kharrazi will not attend the conference because of the statements made by the Jordanian Monarch Abdullah II who has accused Iran of interfering in Iraqi affairs.

n-gram matches: 24 unigrams, 19 bigrams, 15 trigrams, and 12 4-grams
human scores: Adequacy:5,4 Fluency:5,4

Systran v. SMT



Implications for Research

- Higher Bleu score does not guarantee genuine improvement in translation quality
- It is therefore inappropriate and insufficient to:
 - Run workshops to compare systems using Bleu alone
 - Compare systems which employ heterogeneous strategies using Bleu
 - Report translation improvements in conference papers without examples and manual verification
 - Dismiss research which fails to improve Bleu as not improving translation quality

Conclusions

- We have shown:
 - Increasing Bleu is insufficient to guarantee genuine improvements
 - Increasing Bleu is unnecessary to have actual improvements
- Breaks our fundamental assumption that Bleu correlates with human judgments
- Implies that current methodology for evaluation of MT research is flawed
- We must develop a new evaluation methodology

Thank you!

What Should We Do Instead?

- Human evaluation
- Careful experimental design with clear, testable hypothesis
- Focused manual evaluation to see whether it is true
- Show examples in papers
- Publish all translations online

When Can We Use Bleu?

- To compare different versions during system development
- As an objective function for minimum error rate training
- As a “sanity check” prior to doing human evaluation

Other Known Deficiencies of Bleu

- Scores hard to interpret
- Different number of references lead to radically different scores
- Does not work on a per sentence level
- No weight given to content-bearing words