Evaluating translation quality - part 2

Machine Translation
Lecture 10

Instructor: Chris Callison-Burch
TAs: Mitchell Stern, Justin Chiu

Website: mt-class.org/penn

BLEU

- **B**i**L**ingual **E**valuation **U**nderstudy
- Uses multiple reference translations
- Look for n-grams that occur anywhere in the sentence

American plane, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, , Florida

2-gram precision = 10/17

Hyp	**appeared calm** when **he was** taken **to the American plane** , **which will to Miami , Florida .**

n-gram precision

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

1-gram precision = 15/18 = .83
2-gram precision = 10/17 = .59
3-gram precision = 5/16 = .31
4-gram precision = 3/15 = .20

- Geometric average

(0.83 * 0.59 * 0.31 * 0.2)^(1/4) = 0.417
or equivalently
exp(ln .83 + ln .59 + ln .31 + ln .2/4) = 0.417

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Hyp	to the American plane

Better?

Hyp	to the American plane

1-gram precision = 4/4 = 1.0
2-gram precision = 3/3 = 1.0
3-gram precision = 2/2 = 1.0
4-gram precision = 1/1 = 1.0

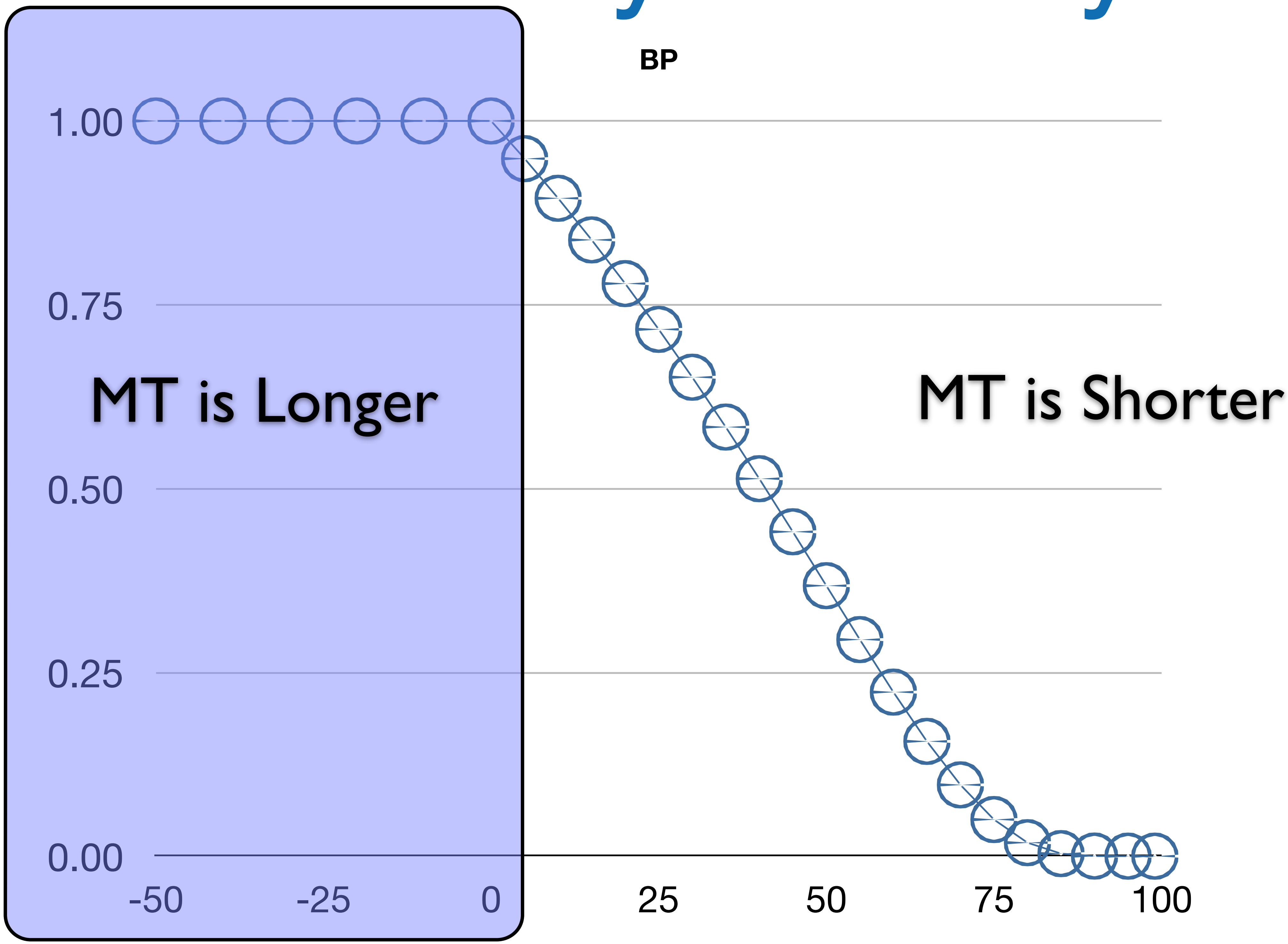
$$\exp(\ln 1 + \ln 1 + \ln 1 + \ln 1) = 1$$

Brevity Penalty

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

- c is the length of the corpus of hypothesis translations
- r is the effective reference corpus length
- The effective reference corpus length is the sum of the single reference translation from each set that is closest to the hypothesis translation.

Brevity Penalty



Difference with effective reference length (%)

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida. $c = 18$

$$BP = \exp(1-(20/18)) = 0.89$$

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	to the American plane $c = 4$

$$BP = \exp(1-(20/4)) = 0.02$$

BLEU

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

- Geometric average of the n-gram precisions
- Optionally weight them with w
- Multiplied by the brevity penalty

BLEU

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

$$\exp(1-(20/18)) * \exp((\ln .83 + \ln .59 + \ln .31 + \ln .2)/4) = \mathbf{0.374}$$

Hyp	to the American plane

$$\exp(1-(20/4)) * \exp((\ln 1 + \ln 1 + \ln 1 + \ln 1)/4) = \mathbf{0.018}$$

Problems with BLEU

- (Discuss with your neighbor)

Problems with BLEU

- **Synonyms and paraphrases** are only handled if they are in the set of multiple reference translations
- The scores for **words are equally weighted** so missing out on content-bearing material brings no additional penalty.
- The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate **recall**.

More Metrics

- WER - word error rate
- PI-WER - position independent WER
- METEOR - **M**etric for **E**valuation of **T**ranslation with **E**xplicit **OR**dering
- TERp - **T**ranslation **E**dit **R**ate **p**lus

Even More Metrics

Metric IDs	Participant
AMBER, AMBER-NL, AMBER-IT	National Research Council Canad
F15, F15G3	Koç University (Bicici and Yuret,
METEOR-1.3-ADQ, METEOR-1.3-RANK	Carnegie Mellon University (Denl
MTERATER, MTERATER-PLUS	Columbia / ETS (Parton et al., 20
MP4IBM1, MPF, WMPF	DFKI (Popović, 2011; Popović et
PARSECONF	DFKI (Avramidis et al., 2011)
ROSE, ROSE-POS	The University of Sheffield (Song
TESLA-B, TESLA-F, TESLA-M	National University of Singapore
TINE	University of Wolverhampton (Ri
BLEU	provided baseline (Papineni et al.,
TER	provided baseline (Snover et al., 2

How do we know which metric is best?

- Measure correlation with human judgments
- How do people evaluation MT quality

Manual Evaluation

Source: Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

Reference: These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ● 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ● 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ● 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	● 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ● 4 ○ 5 Worst

5-point scales

Fluency

How do you judge the fluency of this translation?

5 = Flawless English
4 = Good English
3 = Non-native English
2 = Disfluent English
1 = Incomprehensible

Adequacy

How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?

5 = All
4 = Most
3 = Much
2 = Little
1 = None

Heather Locklear Arrested for driving under the influence of drugs


Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses." Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms.

- Why was Heather Locklear arrested?
  - She was arrested on suspicion of driving under the influence of drugs.
- Why did the bystander call emergency services?
  - He was concerned for Ms. Locklear's life.
- Where did the witness see her acting abnormally?
  - Pulling out of parking in Montecito

Was arrested actress Heather Locklear because of the driving under the effect of an unknown medicine

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse `People'. The female witness told in detail, that Locklear 'pressed `after 16:30 clock accelerator and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when it went backwards, she pulled itself together unites Male at their sunglasses'. A little later the female witness that did probably

- Why was Heather Locklear arrested?

Driving while medicated

- Why did the bystander call emergency services?

There was a lot of noise

- Where did the witness see her acting abnormally?

In a parking lot

. Medikamentes unknown have the effect of a fahrens under actress heather locklear arrested

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

In Santa. One is, melrose place the series of the role of the 'remember the locklear actress the heather this weekend, because of the fahrens Barbara (California) in effect unknown medikamentes arrested People 'magazine. The traffic police California, spokesman for the auszufahren montecito reported in its way from tried parklücke type strange right, you have seen as a witness. . In some Zeitung, as and when they tried to a great deal of 30 p.m., witness the detail of history locklear after 16: that durchdrückte peddle noise and its progress was made parklücke for the car or moving backwards, they had they times of their sonnenbrille ' . The first was probably recognised that locklear a nearby road and anhielt, had not, with the witness to the car off

- Why was Heather Locklear arrested?
- Why did the bystander call emergency services?
- Where did the witness see her acting abnormally?

Heather Locklear Arrested for driving under the influence of drugs

Heather Locklear
Photo by: Santa Barbara County Sheriff's Department

The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving und influence of drugs. A witnes viewed her performing inappropriate maneuvers wh trying to take her car out fro parking in Montecito, as rev to People magazine by a spokesman for the Californian Highway Police. The witnes stated that around 4.30pm M Locklear "hit the accelerator violently, making excessive noise while trying to take her car o from the parking with abrupt and forth maneuvers. While reversing, she passed several times in front of his sunglass Shortly after, the witness, wh a first time, apparently had n recognized the actress, saw Ms.

- Why was Heather Locklear arrested?
  - She was arrested on suspicion of driving under the influence of drugs.

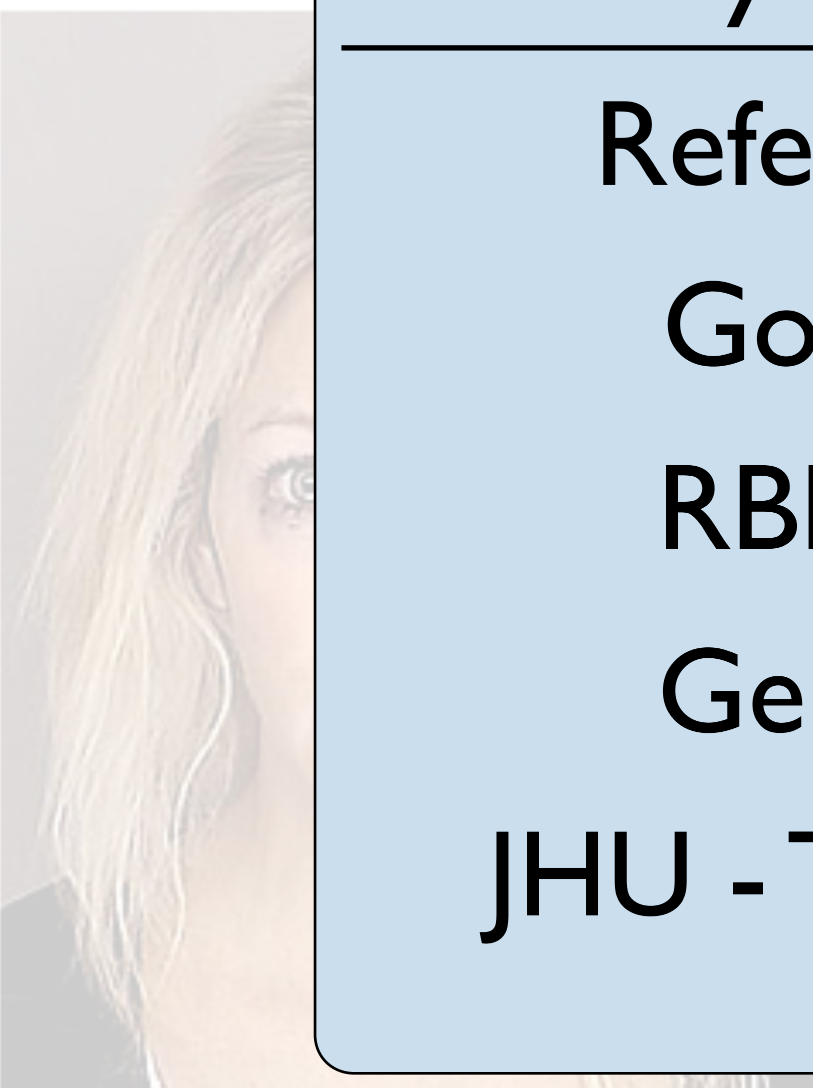
Driving under the influence

Driving while medicated

DUI

Driving while using drugs

Medikamentes



System	Correct Answers
Reference	94%
Google	80%
RBMT5	77%
Geneva	63%
JHU - Tromble	50%

HTER - costs to edit

Reference translation

The man was on assignment from the Ministry of Defense when he left two highly classified documents on a train to Waterloo.

Machine translation

The man was seconded by the Ministry of Defense when he was two extremely confidential documents in a train to Waterloo lost.

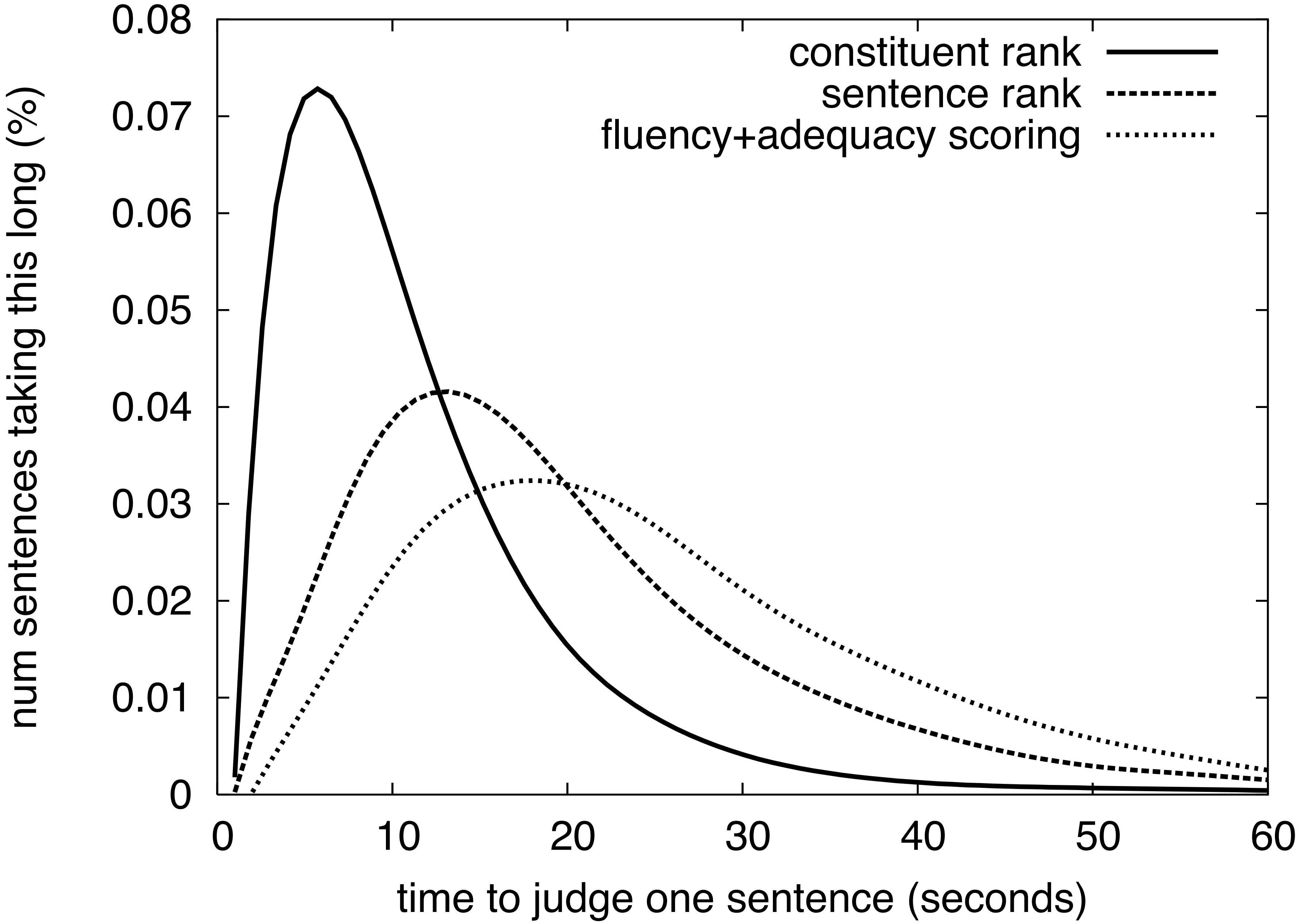
Edited machine translation

The man was **working for** the Ministry of Defense when he **lost** two extremely confidential documents in a train to Waterloo.

Reading Comprehension of Machine Translation

- Jones et al (2005) - Measured translation quality by testing English speakers on a Defense Language Proficiency Test for Arabic
- Read the MT output, and assess how many questions were answered correctly
- Nice, intuitive gauge of how good MT quality actually is

Which type of Human Evaluation is Best?



Using manual judgments to evaluate automatic metrics...

- Measure correlation with human judgments
- System-level correlation
- Sentence-level correlation

Calculating Correlation

- The human evaluation metrics provide a ranking of the systems
  - So do the automatic metrics
- Calculate the correlation between the two lists
  - Metrics with higher correlation better predict human judgments

Spearman's rank correlation coefficient

- For system-level correlation

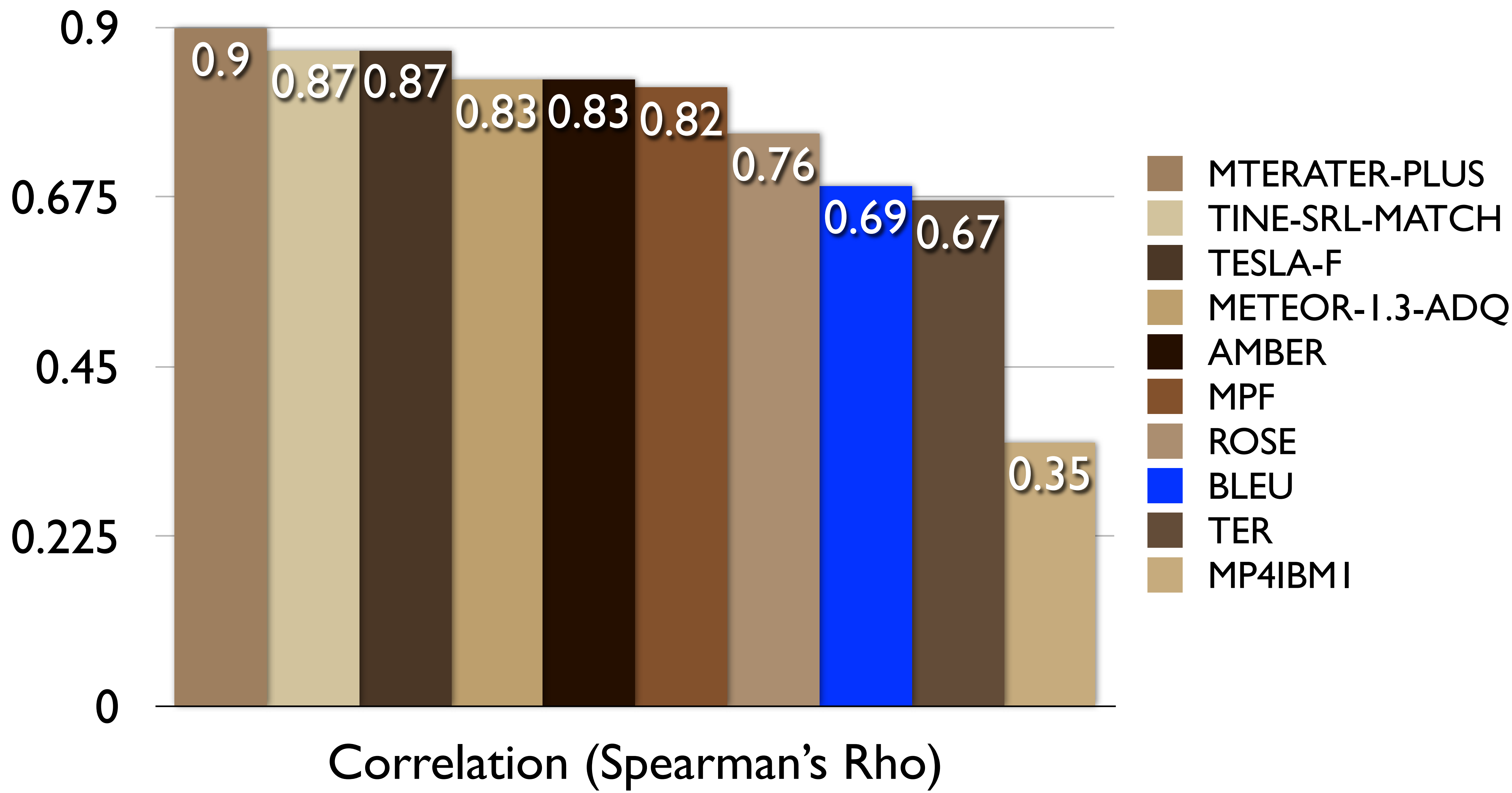
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Kendall's Tau

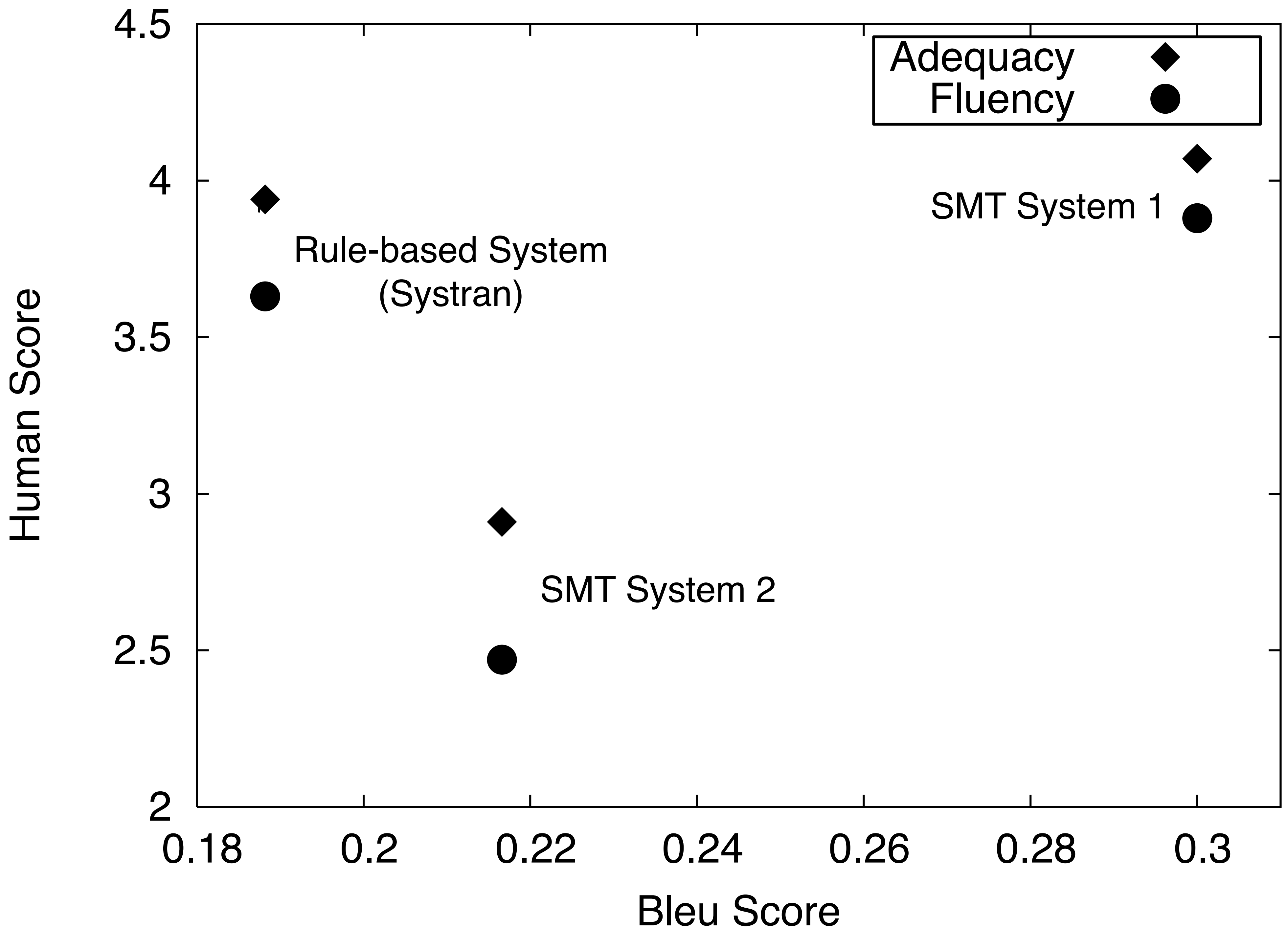
- Segment level evaluation

$$\tau = \frac{\text{num concordant pairs - num discordant pairs}}{\text{total pairs}}$$

Many metrics are better than BLEU



This is bad



Adequacy
Fluency
4.5
4
3.5
3
2.5
2
Human Score
0.18
0.2
0.22
0.24
0.26
0.28
0.3
Bleu Score
Rule-based System
(Systran)
SMT System 1
SMT System 2

Re-evaluating the Role of Bleu in Machine Translation Research

Chris Callison-Burch   Miles Osborne   Philipp Koehn

If Bleu's correlation with human judgments has been overestimated, then the field needs to ask itself whether it should continue to be driven by Bleu to the extent that it currently is. In this paper we give a number of counterexamples for Bleu's correlation with human judgments. We show that under some circumstances an improvement in Bleu is *not sufficient* to reflect a genuine improvement in translation quality, and in other circumstances that it is *not necessary* to improve Bleu in order to achieve a noticeable improvement in translation quality.

Final thoughts on Evaluation

When writing a paper

- If you're writing a paper that claims that
  - one approach to machine translation is better than another, or that
  - some modification you've made to a system has improved translation quality
- Then you need to back up that claim
- Evaluation metrics can help, but good experimental design is also critical

Experimental Design

- Importance of separating out training / test / development sets
- Importance of standardized data sets
- Importance of standardized evaluation metric
- Error analysis
- Statistical significance tests for differences between systems

Evaluation drives MT research

- Metrics can drive the research for the topics that they evaluate
- NIST MT Eval -> DARPA Funding
- Bleu has lead to a focus on phrase-based translation
- Minimum error rate training (next lecture!)
- Other metrics may similarly change the community's focus

Invent your own evaluation metric

- If you think that Bleu is inadequate then invent your own automatic evaluation metric
- Can it be applied automatically?
- Does it correlate better with human judgment?
- Does it give a finer grained analysis of mistakes?

Goals for Automatic Evaluation

- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we're doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score
- Quick to calculate for MERT

Questions?

- Tons of data available at
- http://statmt.org/wmt10/results.html
- http://statmt.org/wmt11/results.html
- http://statmt.org/wmt12/results.html
- http://statmt.org/wmt13/results.html

Reading

- Read 8 from the textbook