NLP Evaluation

CS 585, Fall 2015

Introduction to Natural Language Processing

http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

How to evaluate an NLP system?

How to evaluate an NLP system?

- Many tasks: Classification .. Translation .. etc.

How to evaluate an NLP system?

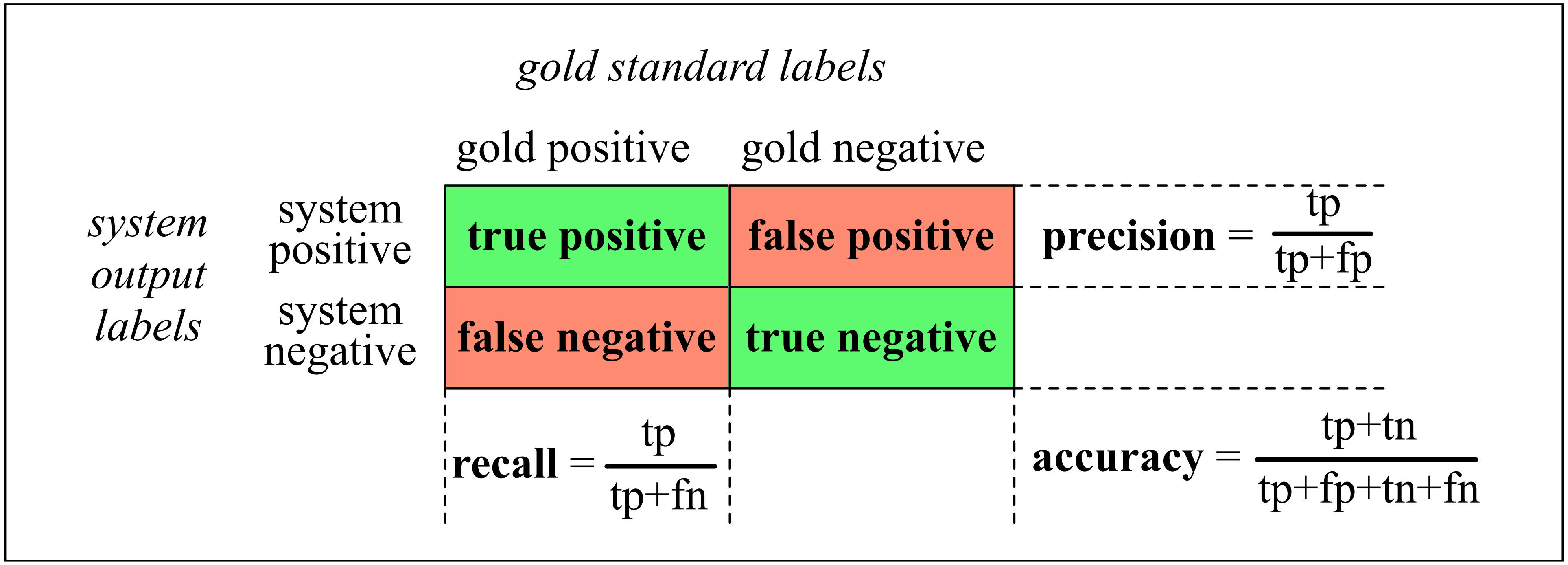
- Many tasks: Classification .. Translation .. etc.
- Extrinsic Evaluation
  Incorporate NLP system into downstream task

How to evaluate an NLP system?

- Many tasks: Classification .. Translation .. etc.
- Extrinsic Evaluation
  Incorporate NLP system into downstream task
- Intrinsic Evaluation
  - Automatic Evaluation
    - Does system agree with pre-judged examples?
  - Human Post-hoc Evaluation

- Questions
  - **What metrics to use?**
  - **How to deal with complex outputs like translations?**
  - Are the human judgments ...
    - ... measuring something real?
    - ... reliable?
  - Is the sample of texts sufficiently representative?
  - How reliable or certain are the results?

Classification metrics

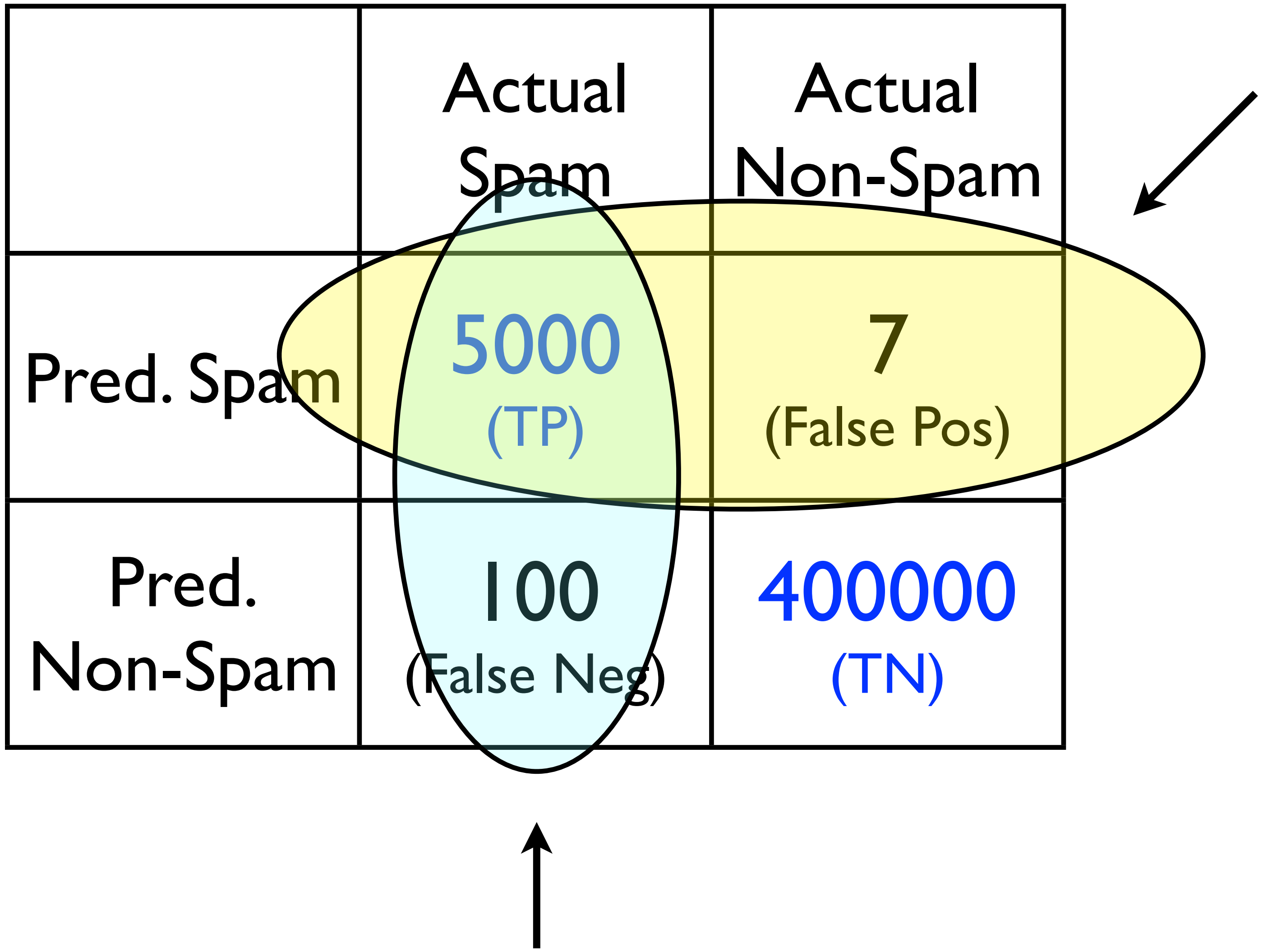


gold standard labels

		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{\text{tp}}{\text{tp+fn}}$		**accuracy** $= \frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Figure 7.4 Contingency table

Confusion matrix



	Actual Spam	Actual Non-Spam
Pred. Spam	5000 (TP)	7 (False Pos)
Pred. Non-Spam	100 (False Neg)	400000 (TN)

Precision = TP / (TP + FP)
= P(correct | predpos)

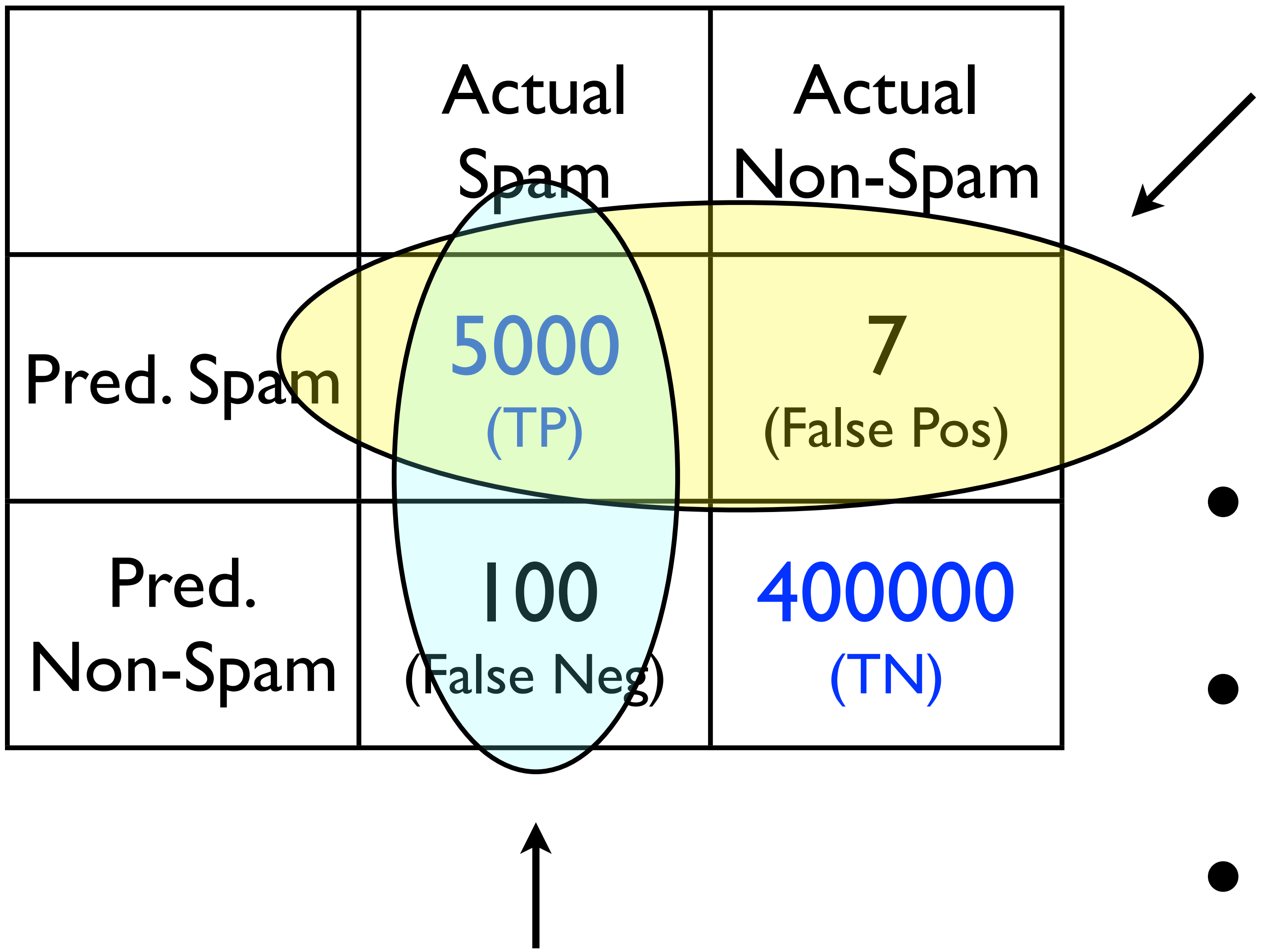
= 5000 / 5007

Recall = TP / (TP + FN)
= P(correct | actualpos)

= 5000 / 5100

http://brenocon.com/confusion_matrix_diagrams.pdf

Confusion matrix



	Actual Spam	Actual Non-Spam
Pred. Spam	5000 (TP)	7 (False Pos)
Pred. Non-Spam	100 (False Neg)	400000 (TN)

Recall = TP / (TP + FN)
= P(correct | actualpos)

= 5000 / 5100

Precision = TP / (TP + FP)
= P(correct | predpos)

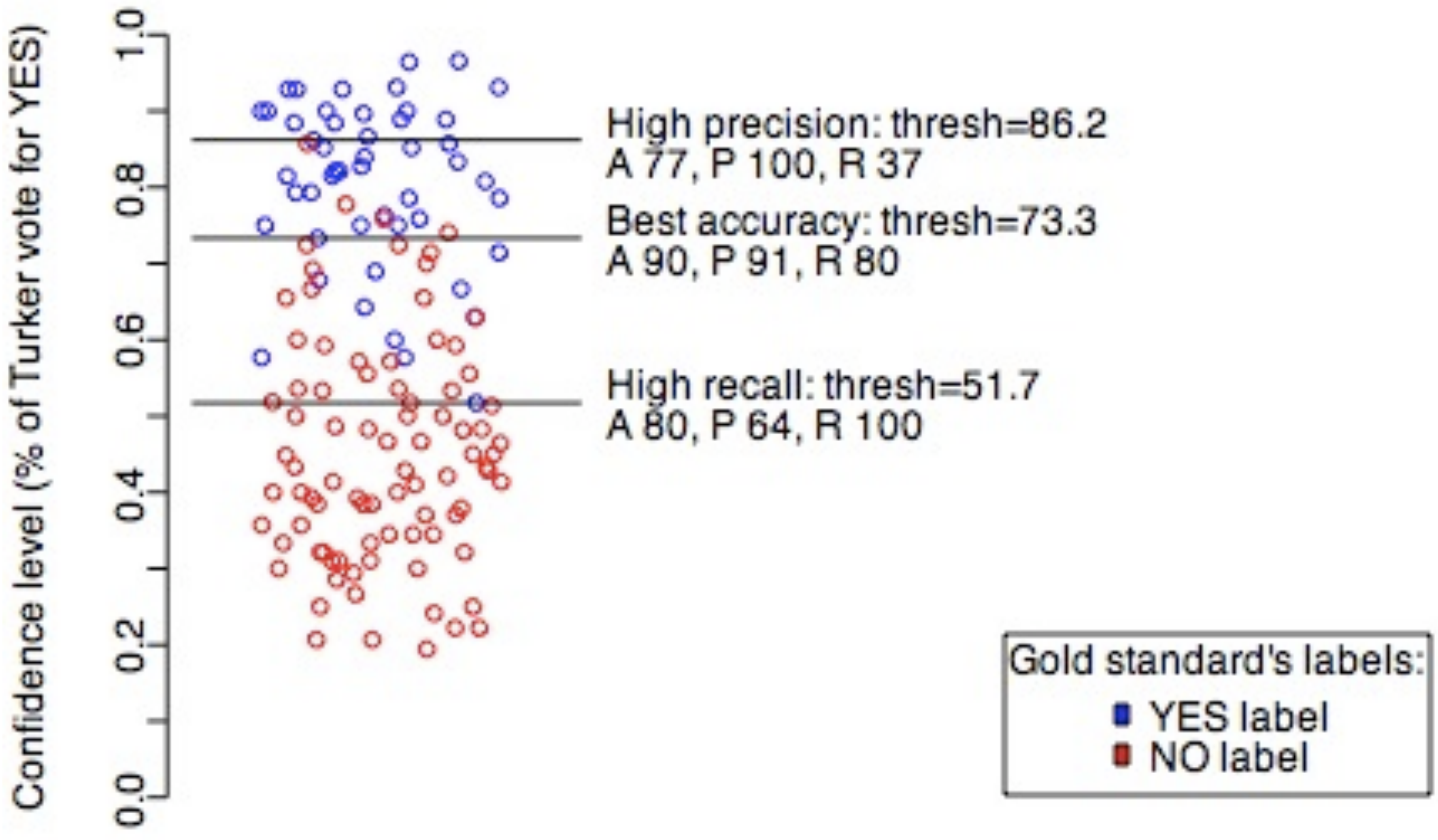
= 5000 / 5007

- You can also just look at the confusion matrix!
- Precision and Recall are metrics for binary classification.
- F-score: harmonic mean of P and R. Cares about getting both moderately high.

http://brenocon.com/confusion_matrix_diagrams.pdf

Trade off Prec vs. Recall

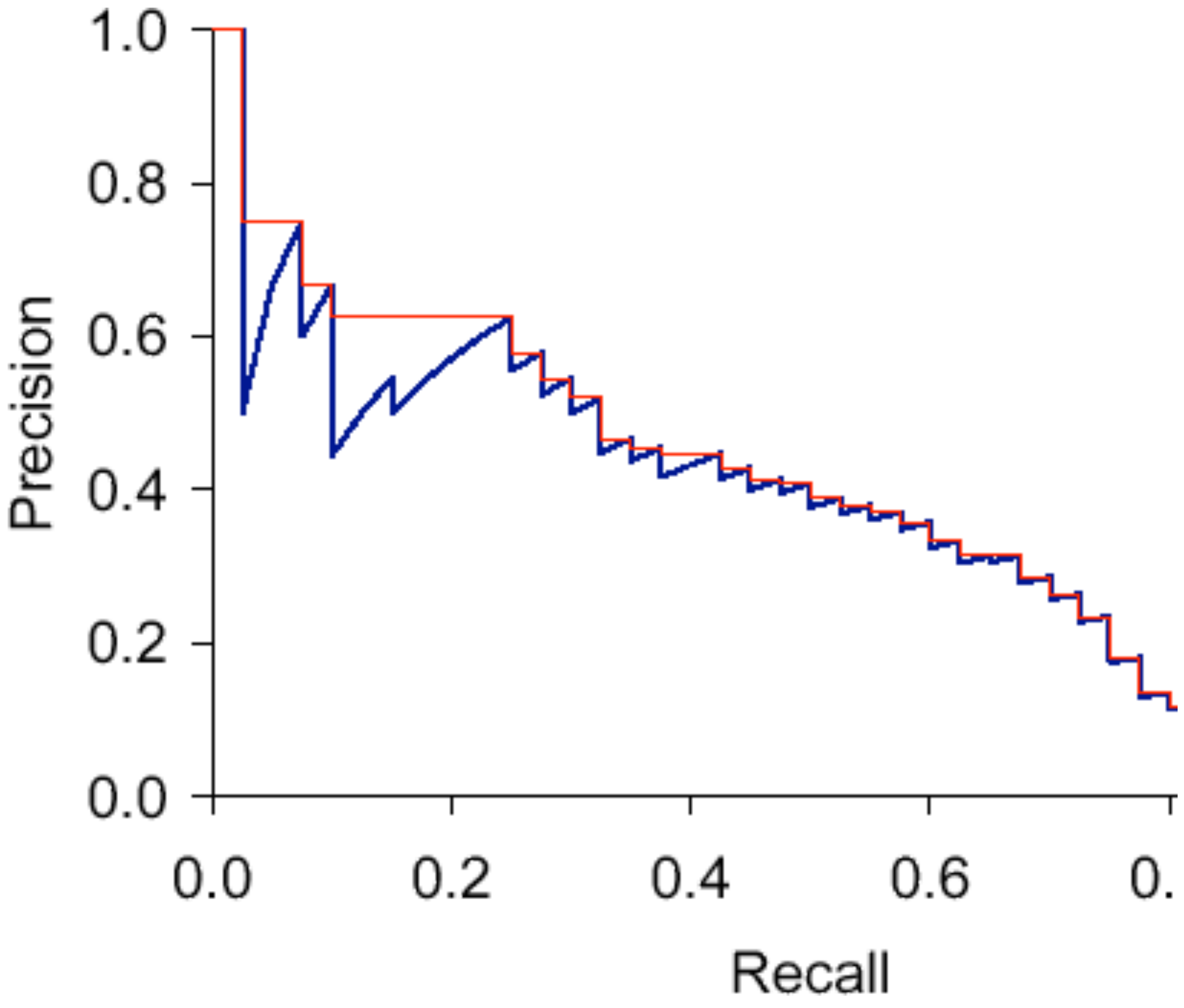
Decide "1" if $p(y = 1|x) > t$ .... could vary threshold ***t***



Above a threshold, classified as Y, below as N
Errors above are false pos; errors below are false neg
Accuracy, Precision, Recall in %
Dots have horizontal jitter (x-axis has no meaning)

http://blog.doloreslabs.com/?p=61

Trade off Prec vs. Recall



MT Evaluation

MT Evaluation

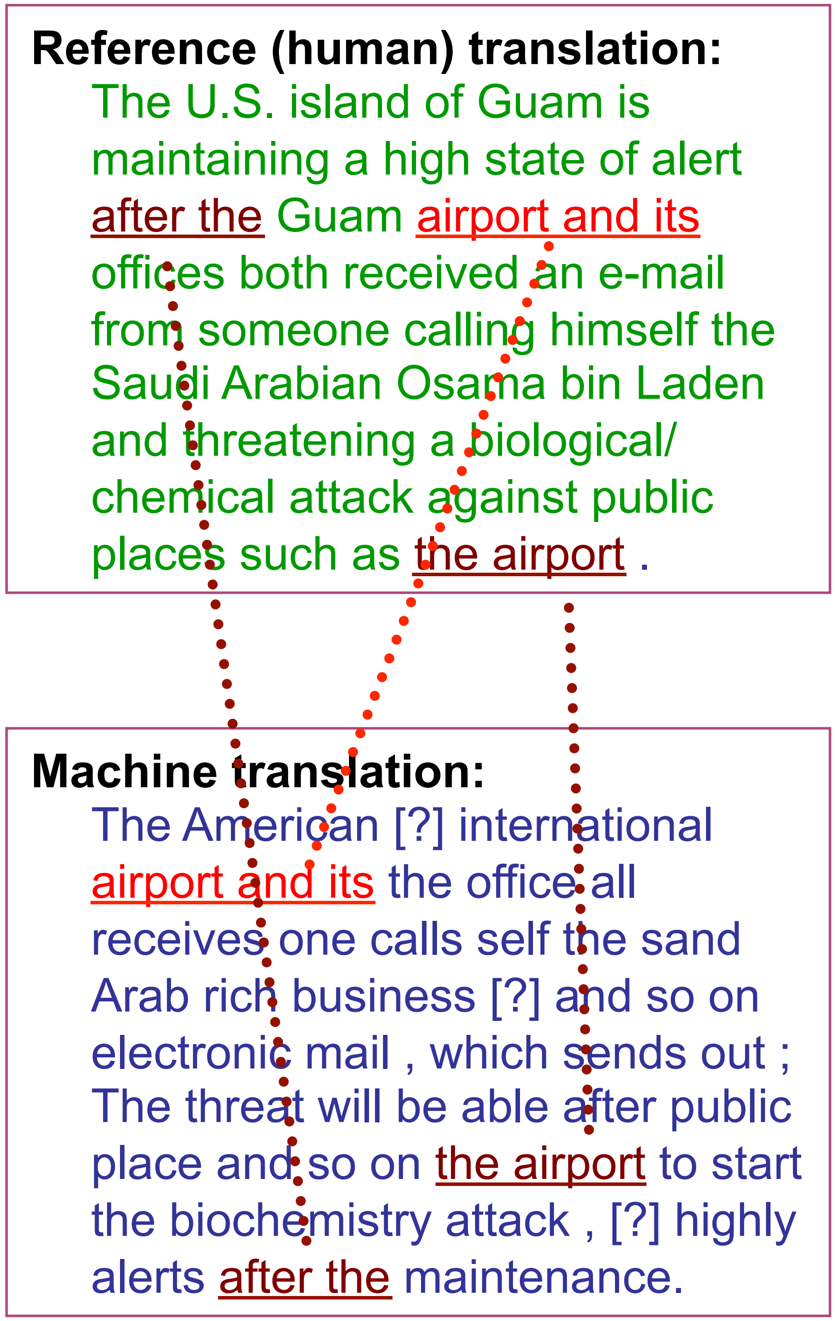
- Manual (the best!?):
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**

- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)

- Automatic metric:
  - WER (word error rate) – why problematic?
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)



Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

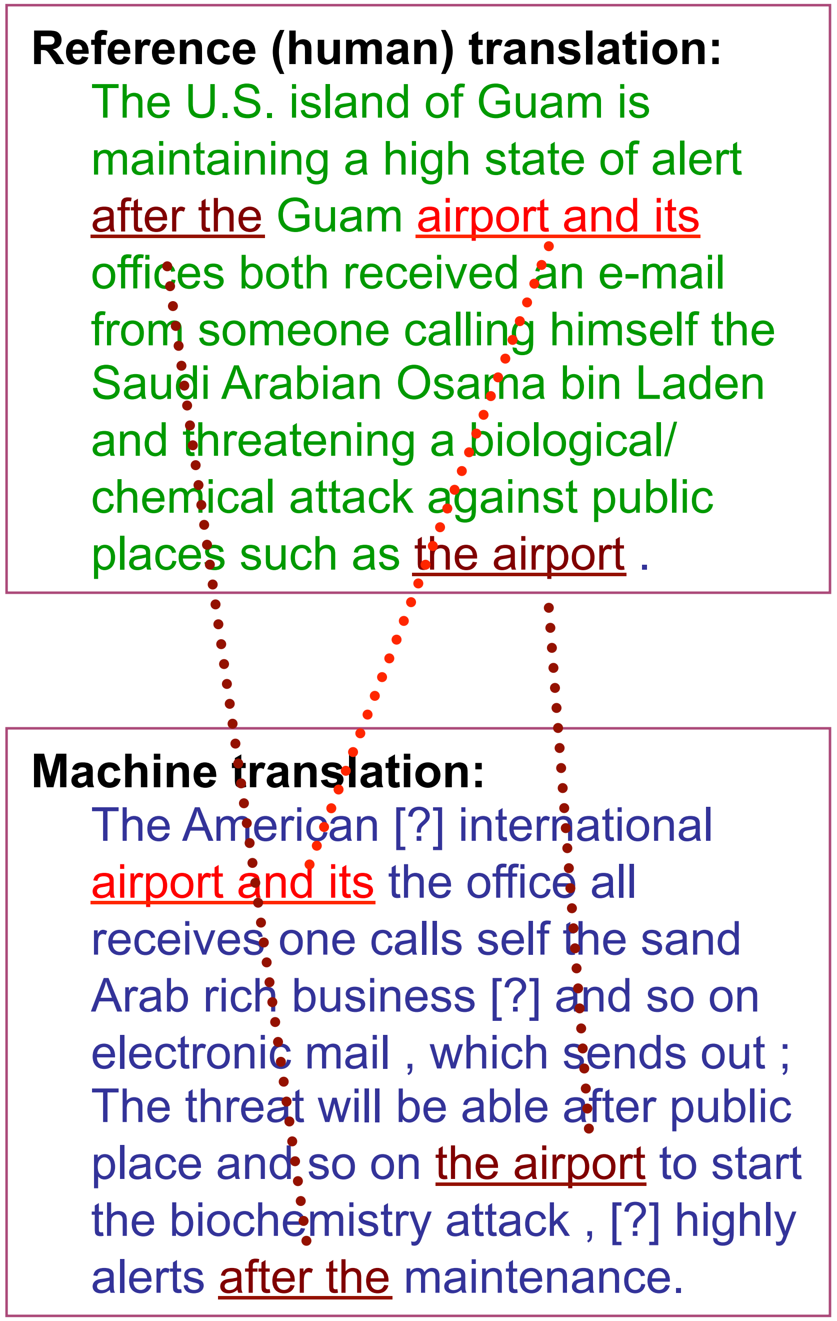
- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words *airport* are only correct if two reference words *airport;* can't cheat by typing out "the the the the the")
  - Do count unigrams also in a bigram for unigram precision, etc.

- Brevity Penalty
  - Can't just type out single word "the" (precision 1.0!)

- It was thought quite hard to "game" the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)



Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

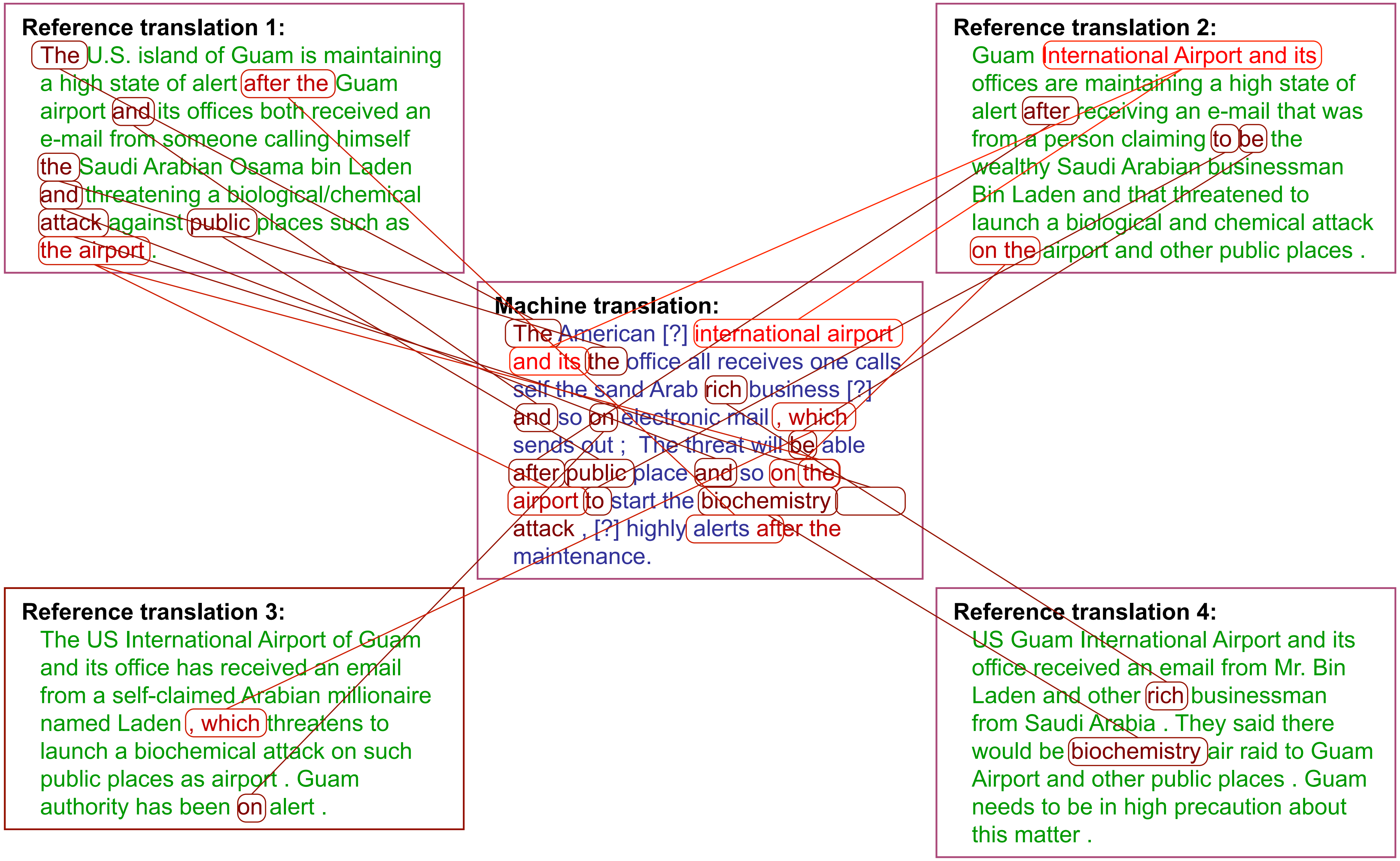
- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it);
there's a smoothed variant for sentence-level

Multiple Reference Translations



Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

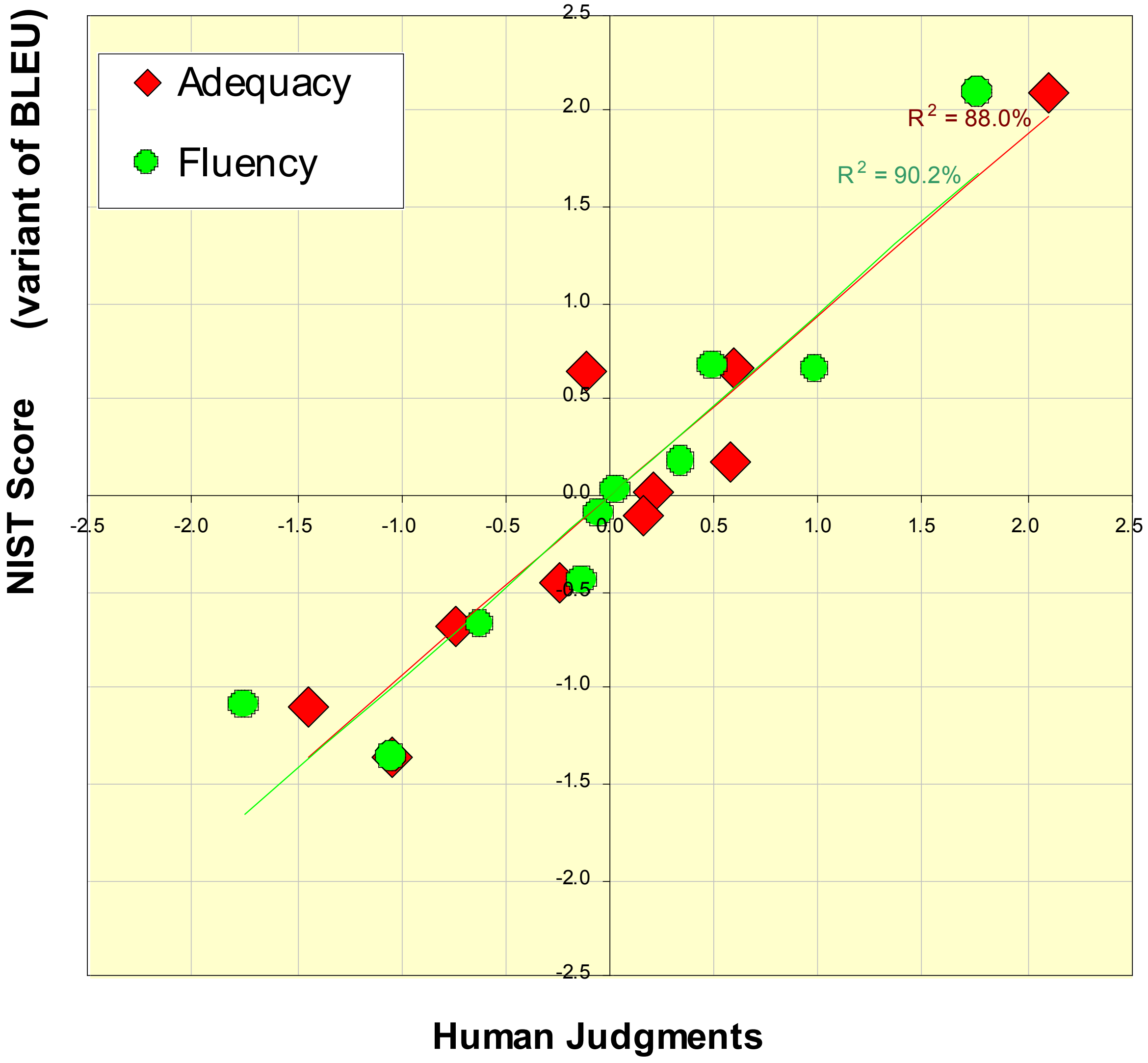
Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

Initial results showed that BLEU predicts human judgments well



slide from G. Doddington (NIST)

- Questions
  - What metrics to use?
  - How to deal with complex outputs like translations?
  - **Are the human judgments ...**
    - **... measuring something real?**
    - **... reliable?**
  - **Is the sample of texts sufficiently representative?**
  - **How reliable or certain are the results?**

Pesky Humans

- Is a task “real”?
- Interannotator agreement rate
  - Accuracy of one human against the other
  - Other metrics: “Cohen’s kappa”
    - normalizes for most-common-baseline issues
- Human performance at task -- upper bound on machine performance?
- What are we trying to measure?
- [EXERCISE]

- stopped here

Significance Testing

- Questions
  - Are the human judgments ...
    - ... measuring something real?
    - ... reliable?
  - **Is the sample of texts sufficiently representative?**
  - How reliable or certain are the results?
  - How to deal with complex outputs like translations?

- Representativeness
  - Is it from the right distribution? Correct domain/genre that we care about?
  - Are there enough examples that we can trust it?

- Representativeness
    - Is it from the right distribution? Correct domain/genre that we care about?
    - Are there enough examples that we can trust it?

- Representativeness
  - Is it from the right distribution? Correct domain/ genre that we care about?
  - Are there enough examples that we can trust it?
- First Q is a judgment call

- Representativeness
  - Is it from the right distribution? Correct domain/genre that we care about?
  - Are there enough examples that we can trust it?

- First Q is a judgment call
- Second Q is a statistical question

Statistical "Significance"

- Assume data was drawn from a greater population.
- If we were to take a new sample, how much would data differ?
  - Or: how much would a *statistic* of that data differ?
  - "Confidence interval"
    (better name: Uncertainty Interval)

Bootstrap test

- [blackboard]

- Inputs
  - Original *data* size N
  - Test statistic: *stat(data)*. e.g.
    - accuracy (numeric)
    - system1 better than system2? (boolean)
- Algorithm
  - For each of 10,000 replications:
    - Draw samp: a sample with replacement from the original data, size N (Many of the original examples will not be in sample)
    - Calculate *stat(samp)*
  - Save all 10,000 *stat(samp)* values. Then analyze
    - Boolean: Calculate proportion that are true
    - Numeric: Calculate mean and standard deviation, and/or plot histogram

Bootstrap test

- 1. Binary null hypothesis (7.2 JM 3ed)
  - p-value: Proportion of replications where the null hypo is true

- 2. Confidence interval (this lecture)
  - Numeric statistic: e.g. accuracy rate
  - The "normal approx" bootstrap CI:
    95% CI = [mean +/- 2*stdev]

Paired tests

- Single dataset. Compare system 1 vs system 2

Power Analysis

- How much data do we have to collect?
- *Power Analysis*: given how big an effect you want to measure, that implies how big N should be
- How to implement
  - Make fake dataset size N, run the bootstrap. Look at whether differences can be detected
    - [IPYNB DEMO]
  - Off-the-shelf formulas, e.g. R *power.t.test()*
  - Rules of thumb: http://www.nrcse.washington.edu/research/struts/chapter2.pdf