SMT Part 7: Evaluation

Laura Jehl (Vertretung),

most slides taken from

http://www.statmt.org/book/

Ten Translations of a Chinese Sentence

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

(a typical example from the 2001 NIST evaluation set)

Evaluation

- How good is a given machine translation system?
- Hard problem, since many different translations acceptable
  → semantic equivalence / similarity
- Evaluation metrics
  - subjective judgments by human evaluators
  - automatic evaluation metrics
  - task-based evaluation, e.g.:
    - how much post-editing effort?
    - does information come across?

Adequacy and Fluency

- Human judgement
  - given: machine translation output
  - given: source and/or reference translation
  - task: assess the quality of the machine translation output

- Metrics

  Adequacy: Does the output convey the same meaning as the input sentence?
  Is part of the message lost, added, or distorted?

  Fluency: Is the output good fluent English?
  This involves both grammatical correctness and idiomatic word choices.

Annotation Tool

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

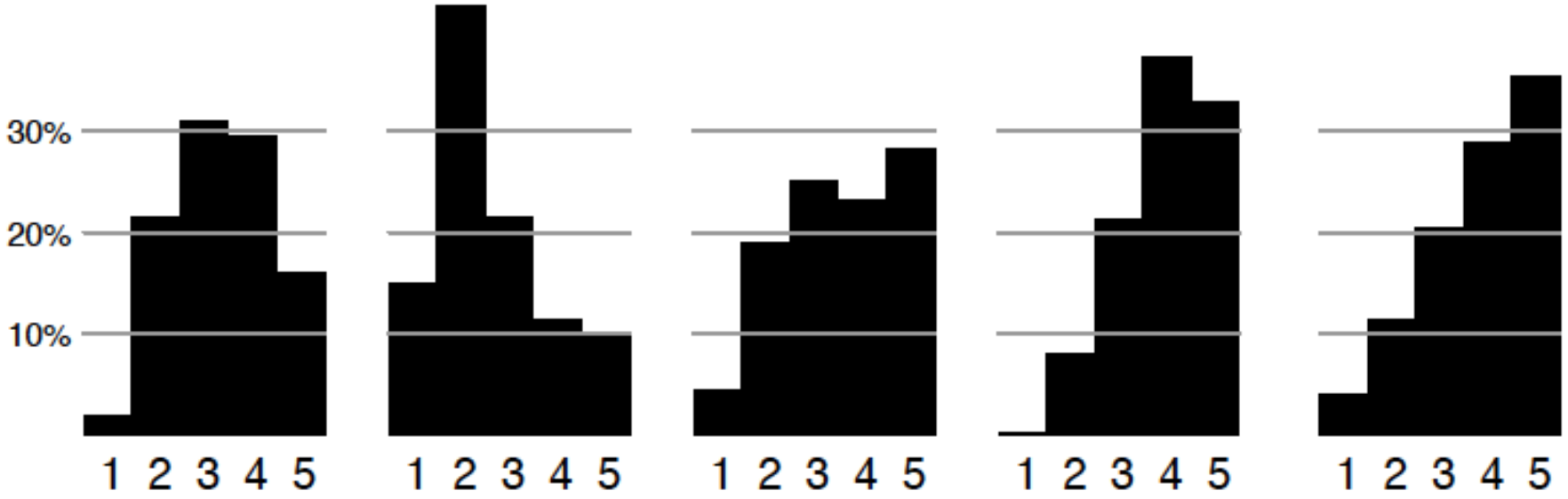
Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Evaluators Disagree

- Histogram of adequacy judgments by different human evaluators



(from WMT 2006 evaluation)

Measuring Agreement between Evaluators

- Kappa coefficient

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

  - $p(A)$: proportion of times that the evaluators agree
  - $p(E)$: proportion of time that they would agree by chance (5-point scale $\rightarrow p(E) = \frac{1}{5}$)

- Example: Inter-evaluator agreement in WMT 2007 evaluation campaign

Evaluation type	$P(A)$	$P(E)$	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226

Ranking Translations

- Task for evaluator: Is translation X better than translation Y? (choices: better, worse, equal)

- Evaluators are more consistent:

Evaluation type	$P(A)$	$P(E)$	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226
Sentence ranking	.582	.333	.373

Хотите светящегося в темноте мороженого? Британский предприниматель создал первое в мире светящееся в темноте мороженое с помощью медузы.

— Source

Fancy a glow-in-the-dark ice cream? A British entrepreneur has created the world's first glow-in-the-dark ice cream - using jellyfish.

— Reference

You do want ice cream luminous in the darkness?

— Translation 1

You want to glowing in the dark ice cream?

— Translation 2

You want the luminous in the dark ice cream?

— Translation 3

Want luminous in the dark ice cream?

— Translation 4

Want to Illuminate the Dark with Ice Cream?

— Translation 5

Goals for Evaluation Metrics

Low cost: reduce time and money spent on carrying out evaluation

Tunable: automatically optimize system performance towards metric

Meaningful: score should give intuitive interpretation of translation quality

Consistent: repeated use of metric should give same results

Correct: metric must rank better systems higher

Human evaluation

- low cost? (✗) – usually turkers or researchers
- tunable? ✗
- meaningful? ✔
- consistent? ✗
- correct? ✔

Other Evaluation Criteria

When deploying systems, considerations go beyond quality of translations

Speed: we prefer faster machine translation systems

Size: fits into memory of available machines (e.g., handheld devices)

Integration: can be integrated into existing workflow

Customization: can be adapted to user's needs

Automatic Evaluation Metrics

- Goal: computer program that computes the quality of translations

- Basic strategy
  - given: machine translation output
  - given: human reference translation
  - task: compute similarity between them

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

- Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

- Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

- F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall

Metric	System A	System B
precision	50%	100%
recall	43%	85%
f-measure	46%	92%

flaw: no penalty for reordering

Word Error Rate

- Minimum number of editing steps to transform output to reference

  match: words match, no cost
  substitution: replace one word with another
  insertion: add word
  deletion: drop word

- Levenshtein distance

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference-length}}$$

Example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

geometric mean of n-gram precisions

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

brevity penalty

- Typically computed over the entire corpus, not single sentences

Example

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Multiple Reference Translations

- To account for variability, use multiple reference translations
  - n-grams may match in any of the references
  - closest reference length used

- Example

SYSTEM: Israeli officials (2-GRAM MATCH) responsibility of (2-GRAM MATCH) airport (1-GRAM) safety

REFERENCES:
Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

METEOR: Flexible Matching

- Partial credit for matching stems

SYSTEM	Jim went home
REFERENCE	Joe goes home

- Partial credit for matching synonyms

SYSTEM	Jim walks home
REFERENCE	Joe goes home

- Use of paraphrases

Critique of Automatic Metrics

- Ignore relevance of words
  (names and core concepts more important than determiners and punctuation)
- Operate on local level
  (do not consider overall grammaticality of the sentence or sentence meaning)
- Scores are meaningless
  (scores very test-set specific, absolute value not informative)
- Human translators score low on BLEU
  (possibly because of higher variability, different word choices)

Automatic evaluation

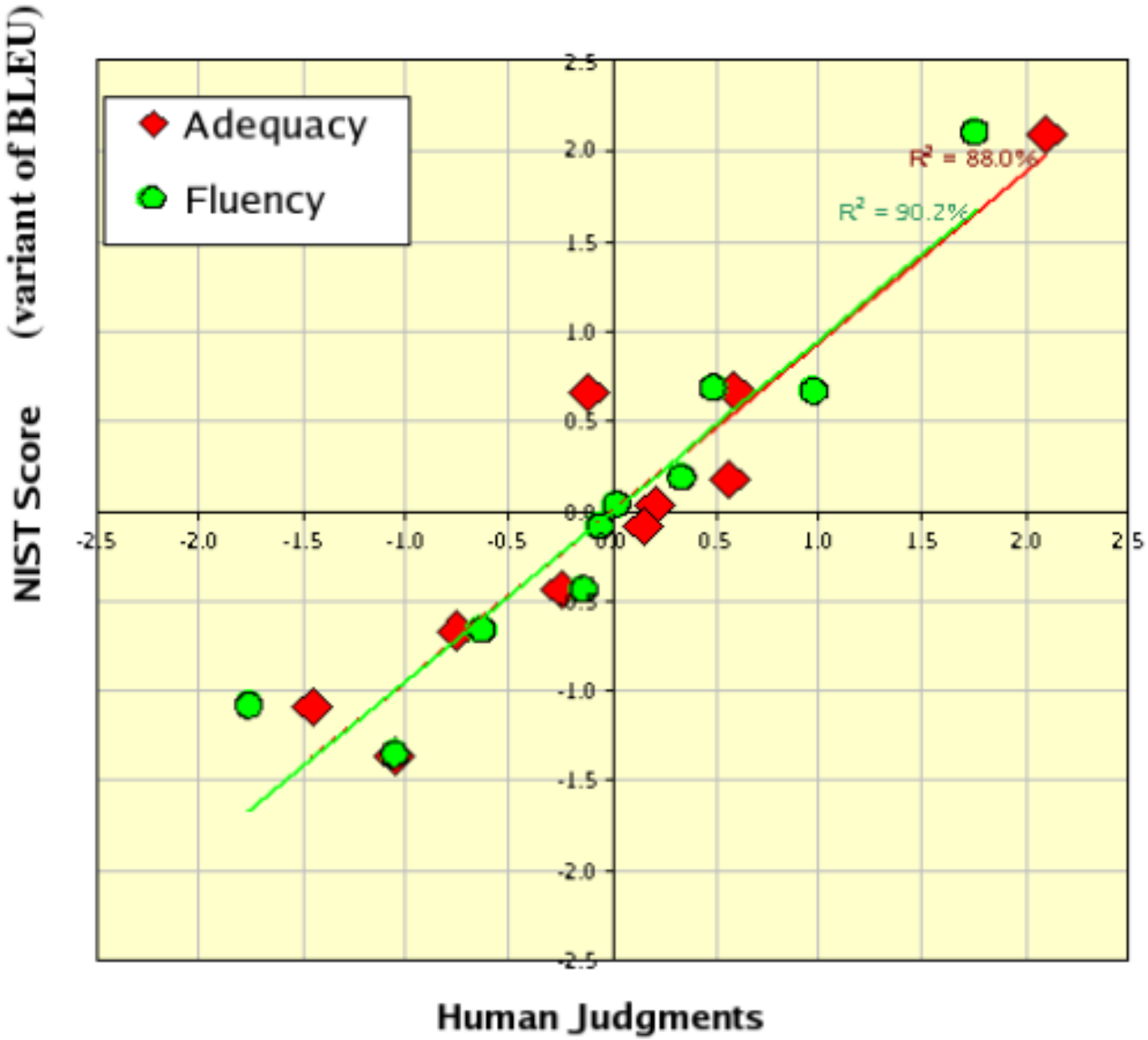
	Human	Automatic
• low cost?	(✗)	✔
• tunable?	✗	✔
• meaningful?	✔	✗
• consistent?	✗	✔
• correct?	✔	(✗)

Evaluation of Evaluation Metrics

- Automatic metrics are low cost, tunable, consistent
- But are they correct?

→ Yes, if they correlate with human judgement

Correlation with Human Judgement



Pearson's Correlation Coefficient

- Two variables: automatic score x, human judgment y
- Multiple systems (x_1, y_1), (x_2, y_2), ...
- Pearson's correlation coefficient r_{xy}:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x\, s_y}$$

- Note:

$$\text{mean } \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\text{variance } s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

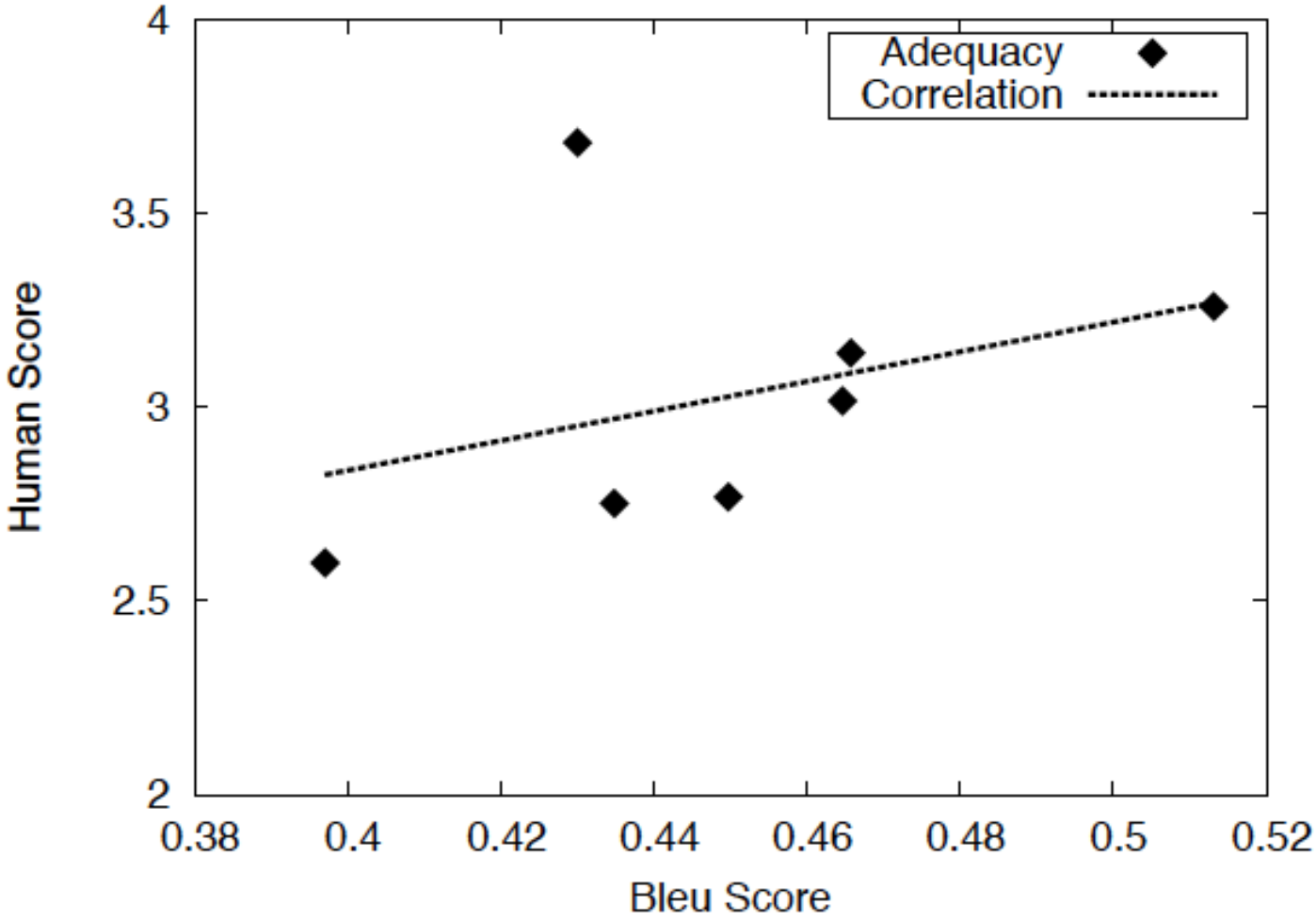
Metric Research

- Active development of new metrics
  - syntactic similarity
  - semantic equivalence or entailment
  - metrics targeted at reordering
  - trainable metrics
  - etc.

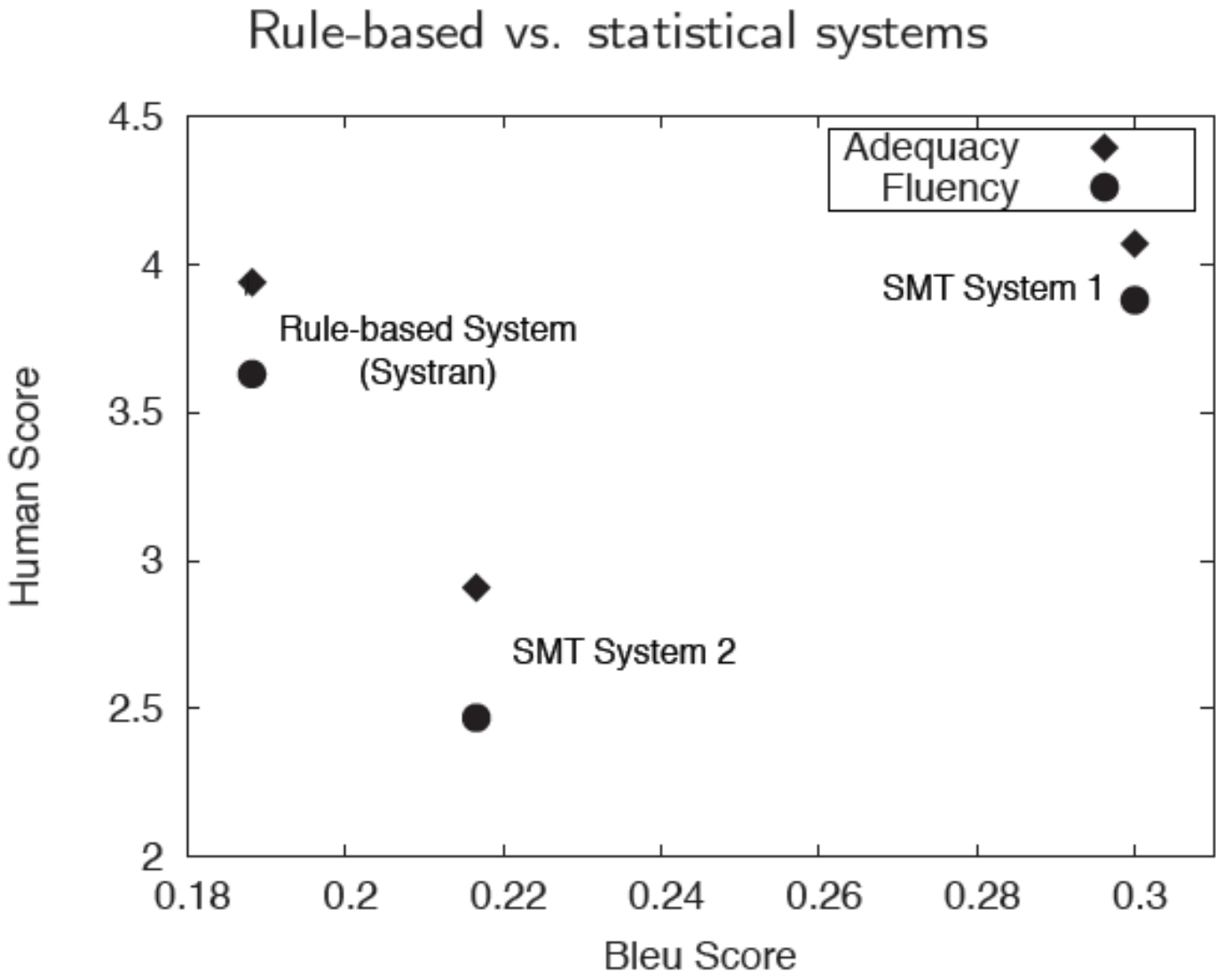
- Evaluation campaigns that rank metrics (using Pearson's correlation coefficient)

Evidence of Shortcomings of Automatic Metrics

Post-edited output vs. statistical systems (NIST 2005)



Evidence of Shortcomings of Automatic Metrics



Automatic Metrics: Conclusions

- Automatic metrics essential tool for system development
- Not fully suited to rank systems of different types
- Evaluation metrics still open challenge

Hypothesis Testing

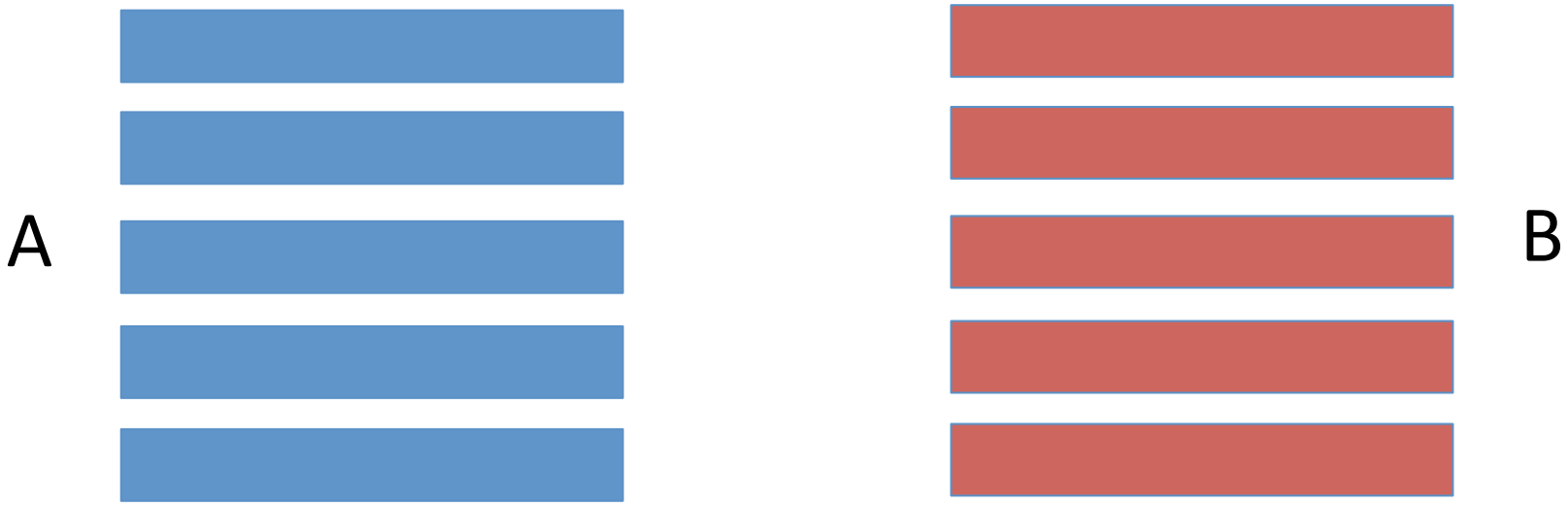
- Situation
  - system A has score x on a test set
  - system B has score y on the same test set
  - $x > y$
- Is system A really better than system B?
- In other words:
  Is the difference in score **statistically significant**?

Core Concepts

- *Null hypothesis*: Assumption that there is no real difference between the systems
- *p-level (p-value)*: probability of seeing the observed or a more extreme result if null-hypothesis is true
  - *p-level < 0.01*: in 99% of cases we expect to see a less extreme result if null-hyp. is true
  - at a *p-level ≤ 0.05* we normally say that there is a significant difference

Testing for significance

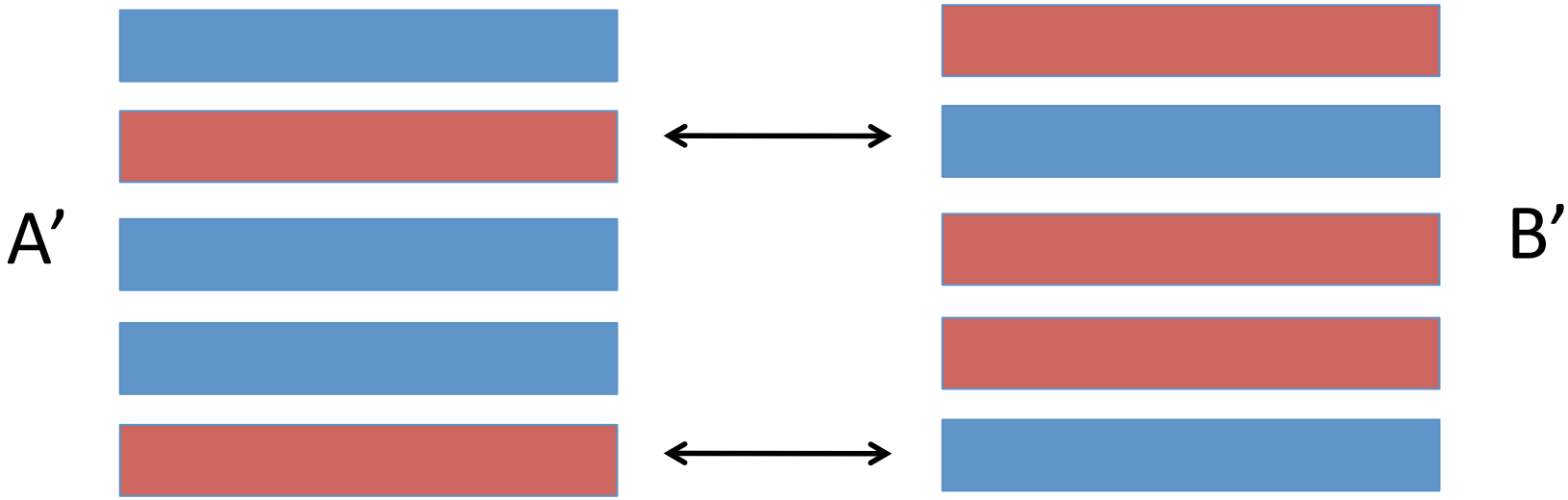
- Idea: If System A and B are not different, then randomly swapping translations between them produces similar scores.



$$|S(A) - S(B)| = 0.6$$

Testing for significance

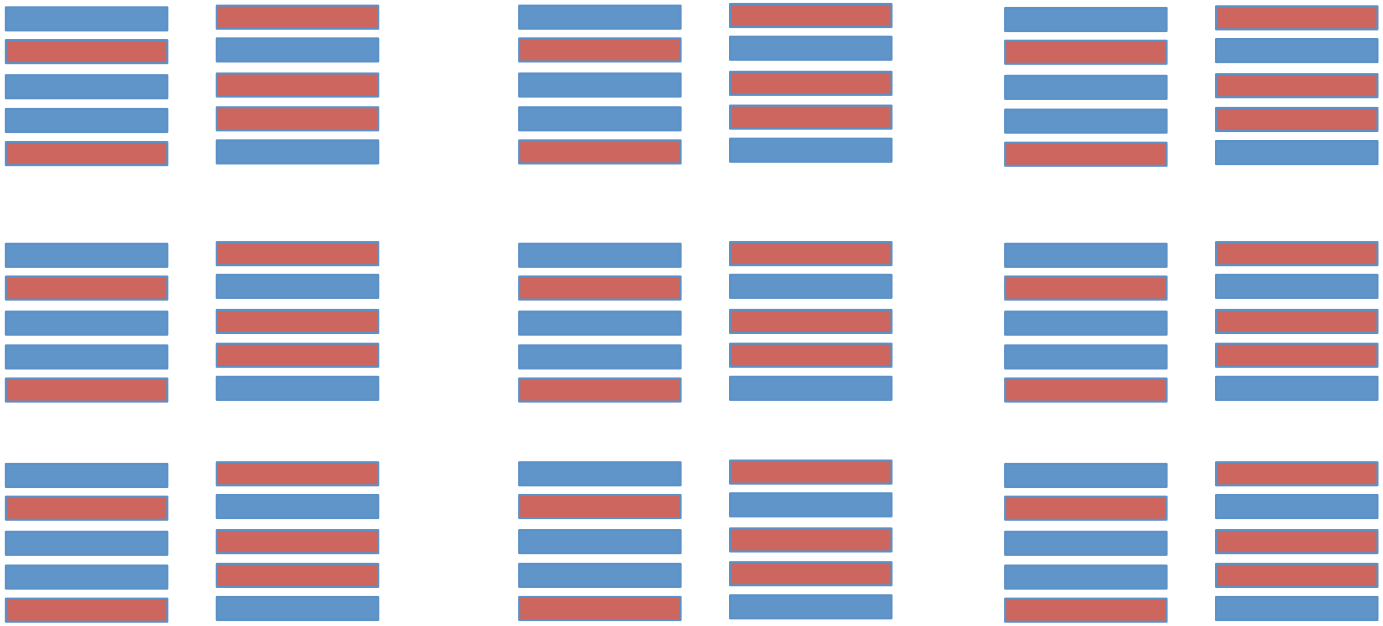
- Idea: If System A and B are not different, then randomly swapping translations between them produces similar scores.



$|S(A') - S(B')| < 0.6$?

Testing for significance

- Repeat this many times and count the number of times that $|S'(A)-S'(B)| > |S(A)-S(B)|$



Testing for significance

- This test is called *Approximate randomization test*
- Usually run for several thousand iterations
- The percentage of times |S(A')-S(B')| > |S(A)-S(B)| is an approximation of the *p-level*
- Rule of thumb: A BLEU difference of 1.0 or more is significant

Approximate Randomization Test:

Set $c = 0$
Compute actual statistic of score differences $|S_X - S_Y|$ on test data for system X, Y
for all random shuffles $r = 0, \ldots, R$ **do**
  for all sentences in test set **do**
    Shuffle variable tuples between system X and Y with probability 0.5
  end for
  Compute pseudo-statistic $|S_{X_r} - S_{Y_r}|$ on shuffled data
  if $|S_{X_r} - S_{Y_r}| \geq |S_X - S_Y|$ **then**
    c ++
  end if
end for
$p = (c + 1)/(R + 1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

source: Stefan Riezler, SMT course notes (2012)

Task-Oriented Evaluation

- Machine translations is a means to an end
- Does machine translation output help accomplish a task?
- Example tasks
  - producing high-quality translations post-editing machine translation
  - information gathering from foreign language sources

Post-Editing Machine Translation

- Measuring time spent on producing translations
  - baseline: translation from scratch
  - post-editing machine translation

  But: time consuming, depend on skills of translator and post-editor

- Metrics inspired by this task
  - TER: based on number of editing steps
    Levenshtein operations (insertion, deletion, substitution) plus movement
  - HTER: manually construct reference translation for output, apply TER
    (very time consuming, used in DARPA GALE program 2005-2011)

Content Understanding Tests

- Given machine translation output, can monolingual target side speaker answer questions about it?

  1. basic facts: who? where? when? names, numbers, and dates
  2. actors and events: relationships, temporal and causal order
  3. nuance and author intent: emphasis and subtext

- Very hard to devise questions

- Sentence editing task (WMT 2009–2010)
  - person A edits the translation to make it fluent (with no access to source or reference)
  - person B checks if edit is correct
    → did person A **understand** the translation correctly?

Games with a purpose

- B: *The trees are on the verge of its greenery or drop in the sky and clouds gather to draw.*

Games with a purpose

- A: *The trees are bare or shortly before throw their leaves and draw storm clouds on the sky.*

Summary

- Machine translation evaluation is hard!
- **Human evaluation** is meaningful and correct, but not tunable or consistent
- Several **automatic evaluation measures** are available which are low-cost, tunable and consistent, but not meaningful
- Correctness of automatic measures can be evaluated by **correlation with human judgments**
- **Significance tests** should be used to determine if two systems are really different

Important concepts

- Adequacy, Fluency
- Kappa-value
- Human vs. automatic evaluation
- BLEU-Score
- Pearson's Correlation
- Approximate randomization test
- Task-based evaluation