Evaluation

Philipp Koehn

17 September 2024

Evaluation

- How good is a given machine translation system?

Evaluation

- How good is a given machine translation system?

- Hard problem, since many different translations acceptable
  $\rightarrow$ semantic equivalence / similarity

Evaluation

- How good is a given machine translation system?

- Hard problem, since many different translations acceptable
  $\rightarrow$ 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?

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)

adequacy and fluency

Adequacy and Fluency

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

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.

Fluency and Adequacy: Scales

Adequacy	
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

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

Hands On: Judge Translations

- Rank according to adequacy and fluency on a 1-5 scale (5 is best)
  - Source:
    L'affaire NSA souligne l'absence totale de débat sur le renseignement
  - Reference:
    NSA Affair Emphasizes Complete Lack of Debate on Intelligence
  - System1:
    The NSA case underscores the total lack of debate on intelligence
  - System2:
    The case highlights the NSA total absence of debate on intelligence
  - System3:
    The matter NSA underlines the total absence of debates on the piece of information

Hands On: Judge Translations

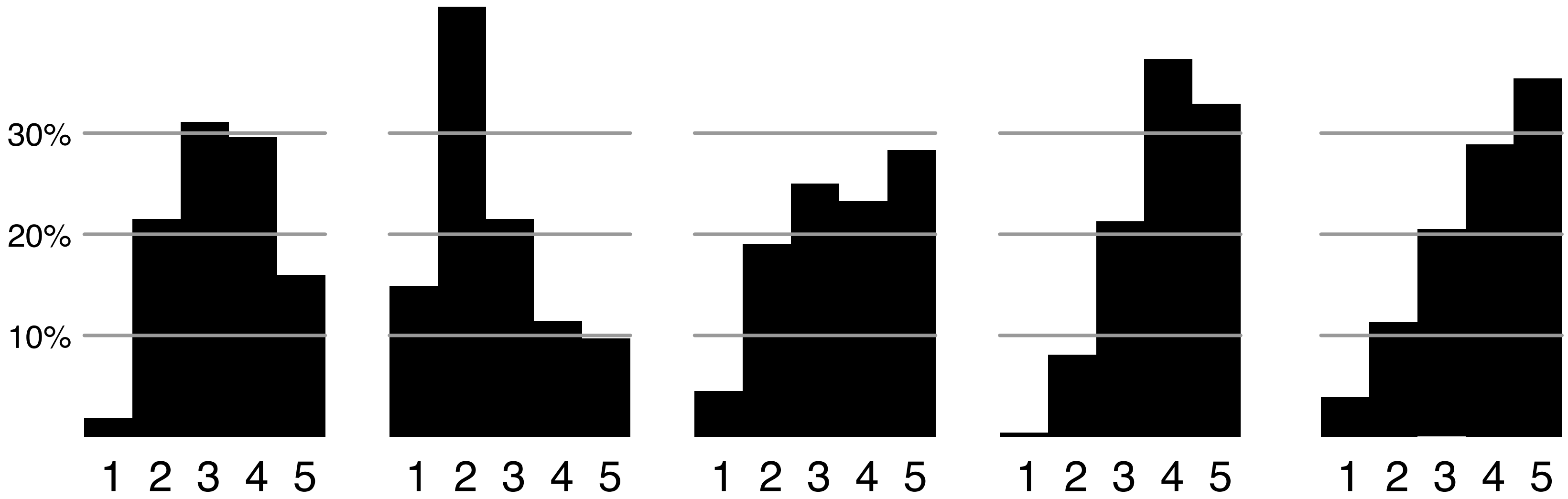
- Rank according to adequacy and fluency on a 1-5 scale (5 is best)
  - Source:
    N'y aurait-il pas comme une vague hypocrisie de votre part ?
  - Reference:
    Is there not an element of hypocrisy on your part?
  - System1:
    Would it not as a wave of hypocrisy on your part?
  - System2:
    Is there would be no hypocrisy like a wave of your hand?
  - System3:
    Is there not as a wave of hypocrisy from you?

Hands On: Judge Translations

- Rank according to adequacy and fluency on a 1-5 scale (5 is best)
  - Source:
    La France a-t-elle bénéficié d'informations fournies par la NSA concernant des opérations terroristes visant nos intérêts ?
  - Reference:
    Has France benefited from the intelligence supplied by the NSA concerning terrorist operations against our interests?
  - System1:
    France has benefited from information supplied by the NSA on terrorist operations against our interests?
  - System2:
    Has the France received information from the NSA regarding terrorist operations aimed our interests?
  - System3:
    Did France profit from furnished information by the NSA concerning of the terrorist operations aiming our interests?

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

Ways to Improve Consistency

- Evaluate fluency and adequacy separately
- Normalize scores
  - use 100-point scale with ”analog” ruler
  - normalize mean and variance of evaluators
- Check for bad evaluators (e.g., when using Amazon Turk)
  - repeat items
  - include reference
  - include artificially degraded translations

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

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 metrics

Automatic Evaluation Metrics

- Goal: computer program that computes the quality of translations

- Advantages: low cost, tunable, consistent

- 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

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

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

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

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)

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

- Typically computed over the entire corpus, not single sentences

Example

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

REFERENCE: Israeli officials are responsible for airport security

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

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: $\boxed{\text{Israeli officials}}$ $\boxed{\text{responsibility of}}$ $\boxed{\text{airport}}$ safety

2-GRAM MATCH | 2-GRAM MATCH | 1-GRAM

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

Translation Edit Rate (TER)

- Account for moves, edit distance with mismatch/insertion/deletion

DELETION MISMATCH MISMATCH MISMATCH

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

REFERENCE: Israeli officials are responsible for airport security

MOVE

SYSTEM B: airport security ~~for~~ Israeli officials are responsible

DELETION

Metric	System A	System B
mismatch	3	0
deletion	1	1
insertion	0	0
move	0	1
total	4	2
TER	4/7	2/7

chrF++

- chrF: Character n-gram F-score (e.g., 6-grams)
- Some nice properties
  - partial credit for morphological variants
  - more credit for longer (content) words than for shorter (function) words
- chrF++: also add F-measure on words and word bigrams to the scoring

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)

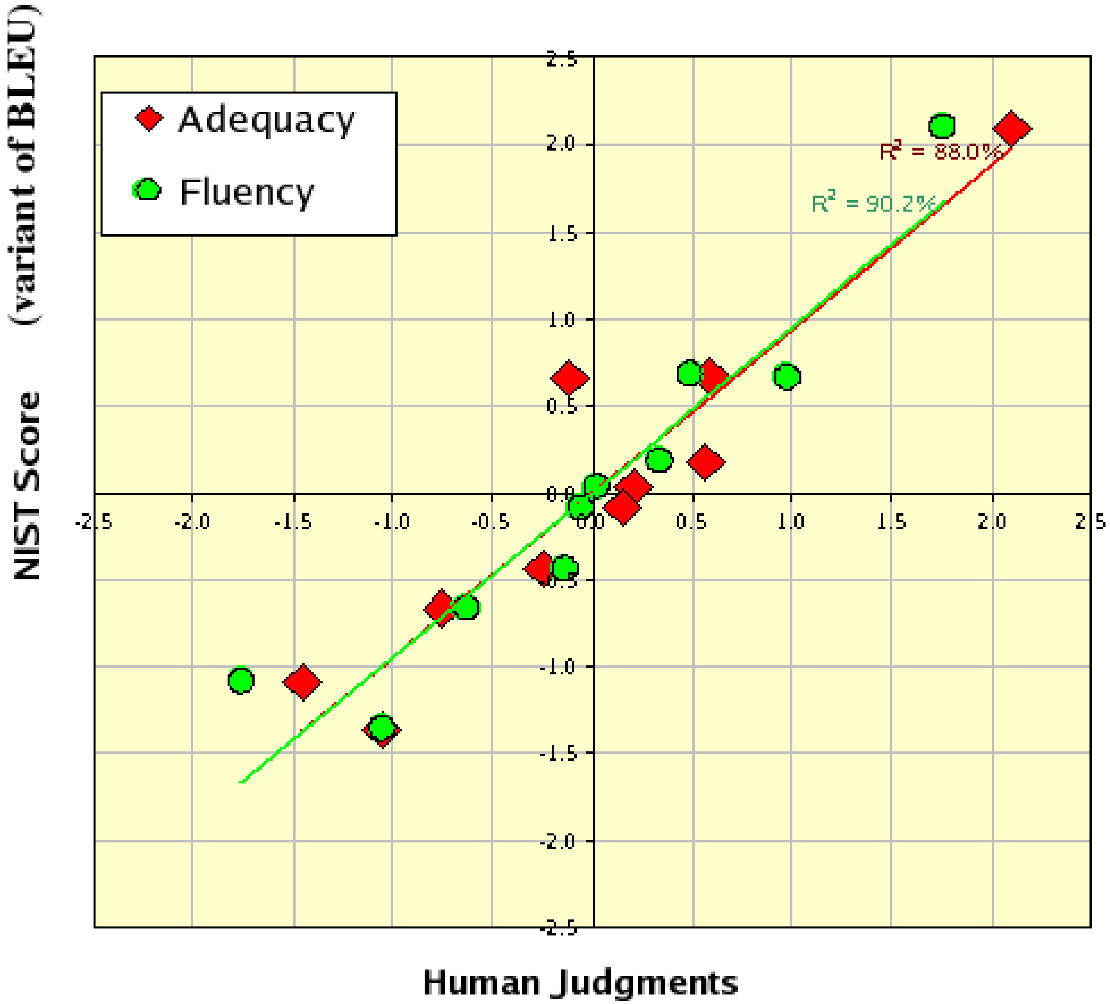
- Human translators score low on BLEU
  (possibly because of higher variability, different word choices)

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}:

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

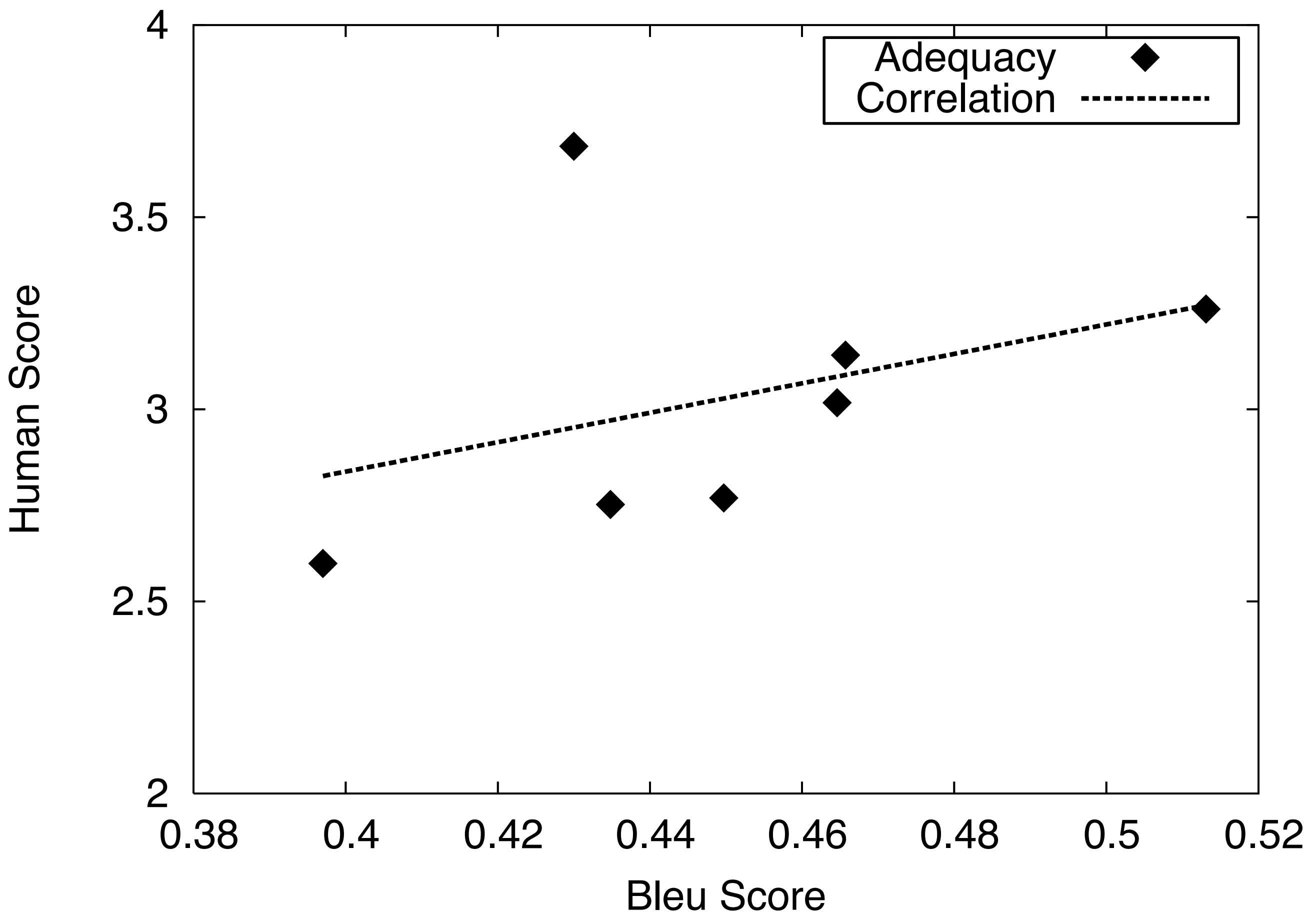
- Note:

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

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

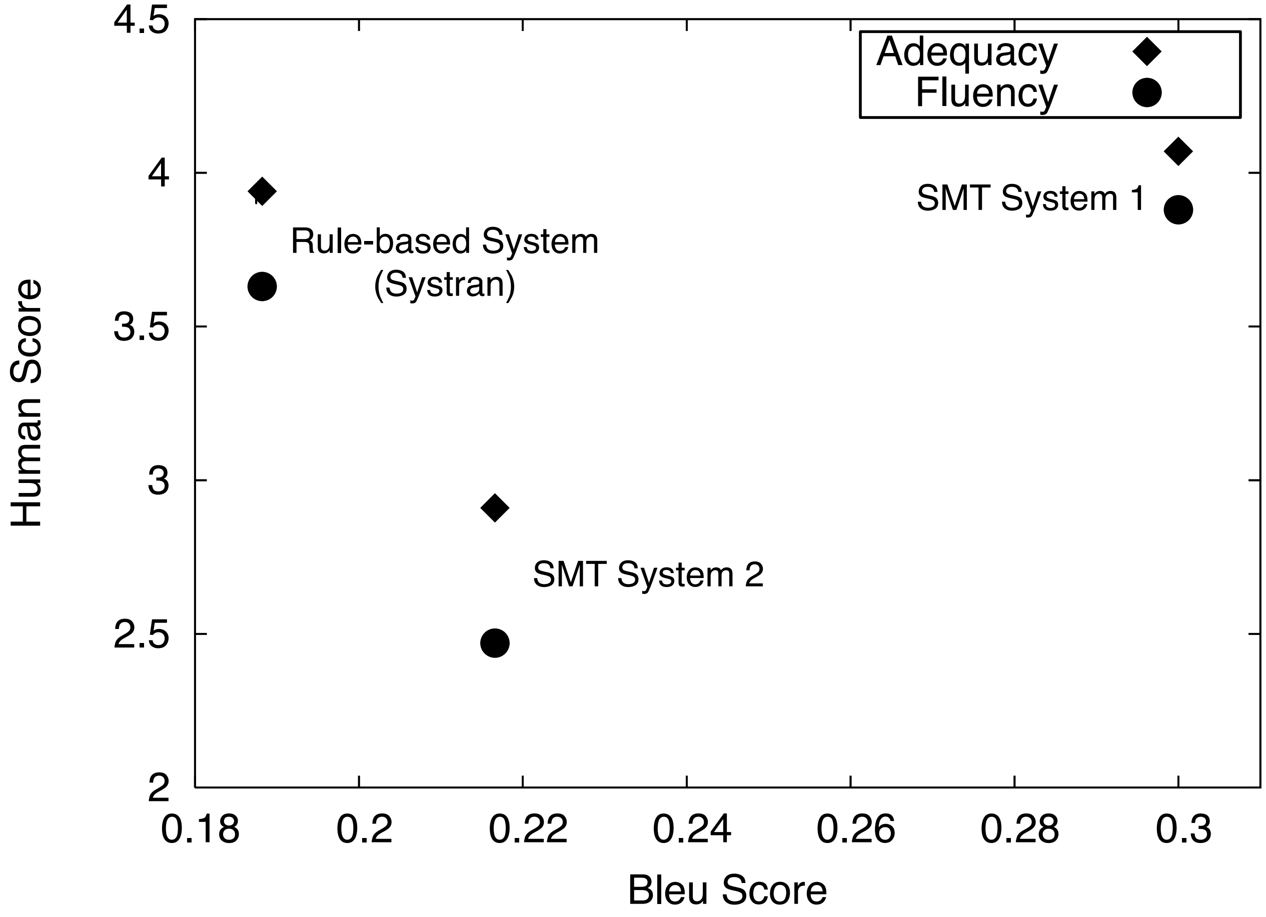
Evidence of Shortcomings of Automatic Metrics

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



Evidence of Shortcomings of Automatic Metrics

Rule-based vs. statistical systems



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)

WMT Metrics Shared Task

- Annual event to evaluate metrics
- Piggy-backs on the WMT General Translation Task
  - new test set every year
  - research systems and commercial systems
  - lately also large language models
  - human evaluation of automatic evaluations
- New metrics proposed
- Evaluation by correlation with human judgments

(WMT 2023)

Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767
Calibri-COMET22	4	0.767
Calibri-COMET22-QE*	4	0.755
YiSi-1	4	0.754
MS-COMET-QE-22*	5	0.744
prismRef	5	0.744
mre-score-labse-regular	5	0.743
BERTscore	5	0.742
XLsim	6	0.719
f200spBLEU	7	0.704
MEE4	7	0.704
tokengram_F	7	0.703
embed_llama	7	0.701
BLEU	7	0.696
chrF	7	0.694
eBLEU	7	0.692
Random-sysname*	8	0.529
prismSrc*	9	0.455

Trained Metrics: COMET

- Two decades of evaluation campaigns for machine translation metrics
  $\rightarrow$ a lot of human judgment data

- Goal: automatic metric that correlates with human judgment

- Make it a machine learning problem
  - input: machine translation, reference translation
  - output: human annotation score

- COMET: Trained neural model for evaluation

Reference-Free Evaluation

- We have data in the form

  input, translation, human reference → human judgment

- We can also train a model on

  input, translation → human judgment

- CometKiwi: trained evaluation model without references
- Also called **quality estimation** or **confidence estimation**

WMT 2024 English-Japanese

Automatic Evaluation

System Name	AutoRank ↓	MetricX ↓	CometKiwi ↑
Unbabel-Tower70B	1.0	2.0	0.762
ONLINE-B	1.4	2.4	0.750
Claude-3.5	1.5	2.3	0.744
Gemini-1.5-Pro	1.7	2.5	0.734
GPT-4	1.7	2.7	0.740
Team-J	1.9	2.9	0.740
NTTSU	1.9	2.6	0.731
CommandR-plus	1.9	2.7	0.730
IOL-Research	2.3	3.1	0.724
Aya23	2.3	3.1	0.719
Llama3-70B §	2.6	3.5	0.714
DLUT-GTCOM	2.6	3.0	0.697
Phi-3-Medium §	2.8	3.6	0.709
ONLINE-W	2.9	3.6	0.700
Mistral-Large §	2.9	3.8	0.707
ONLINE-A	3.0	3.6	0.699
IKUN	3.1	3.7	0.696
IKUN-C	3.9	4.3	0.669
ONLINE-G	6.4	6.6	0.599
AIST-AIRC	6.6	6.5	0.583
UvA-MT	6.7	6.7	0.589
NVIDIA-NeMo †	6.9	6.9	0.582
CycleL	24.0	22.4	0.101

Human Evaluation

Rank	System	Human	AutoRank
1-7	GPT-4	91.0	1.7
1-6	ONLINE-B	90.9	1.4
1-6	HUMAN-A	90.5	-
1-6	CommandR-plus	90.4	1.9
1-6	Claude-3.5	90.1	1.5
7-11	Team-J	89.9	1.9
6-11	Unbabel-Tower70B	89.7	1.0
1-7	Gemini-1.5-Pro	89.4	1.7
6-9	NTTSU	89.3	1.9
8-11	IOL-Research	89.2	2.3
8-11	Aya23	88.6	2.3
12-12	Llama3-70B §	85.7	2.6
13-13	IKUN-C	81.3	3.9

Some Finer Points

- Metrics that match words prefer literal translations

Some Finer Points

- Metrics that match words prefer literal translations

- It matters if the data was created in the same translation direction

  native sentence $\rightarrow$ human translation

  can be used to evaluate

  human translation $\rightarrow$ native sentence

  Translationese: translation shows properties of source language, is not native

Some Finer Points

- Metrics that match words prefer literal translations

- It matters if the data was created in the same translation direction

  native sentence → human translation

  can be used to evaluate

  human translation → native sentence

  Translationese: translation shows properties of source language, is not native

- Quality of human reference translation matters
  (e.g., real problem if translation was created by post-editing output from a specific machine translation model)

Automatic Metrics: Conclusions

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

task-oriented evaluation

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

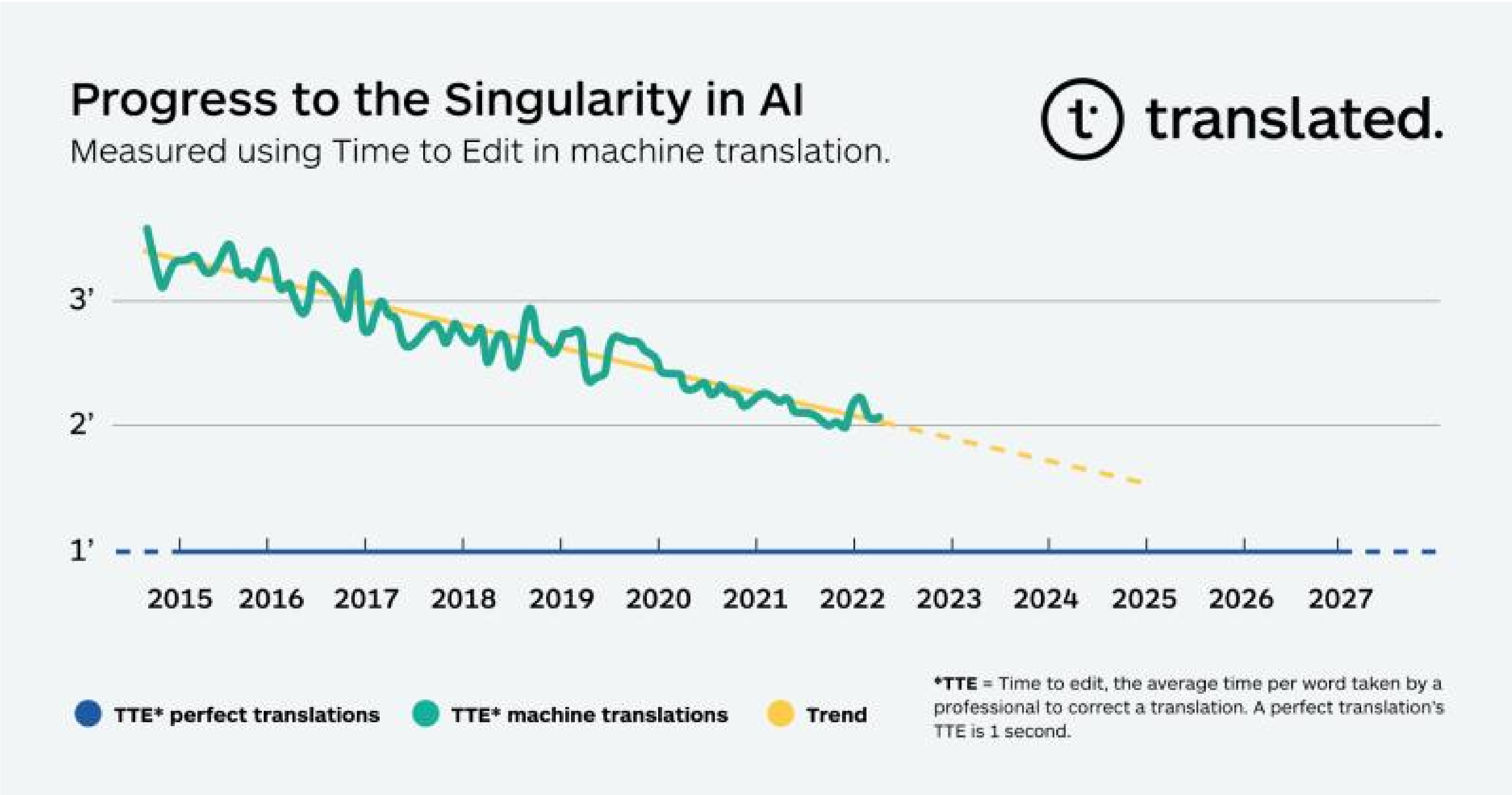
Post-Editing Machine Translation

- 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

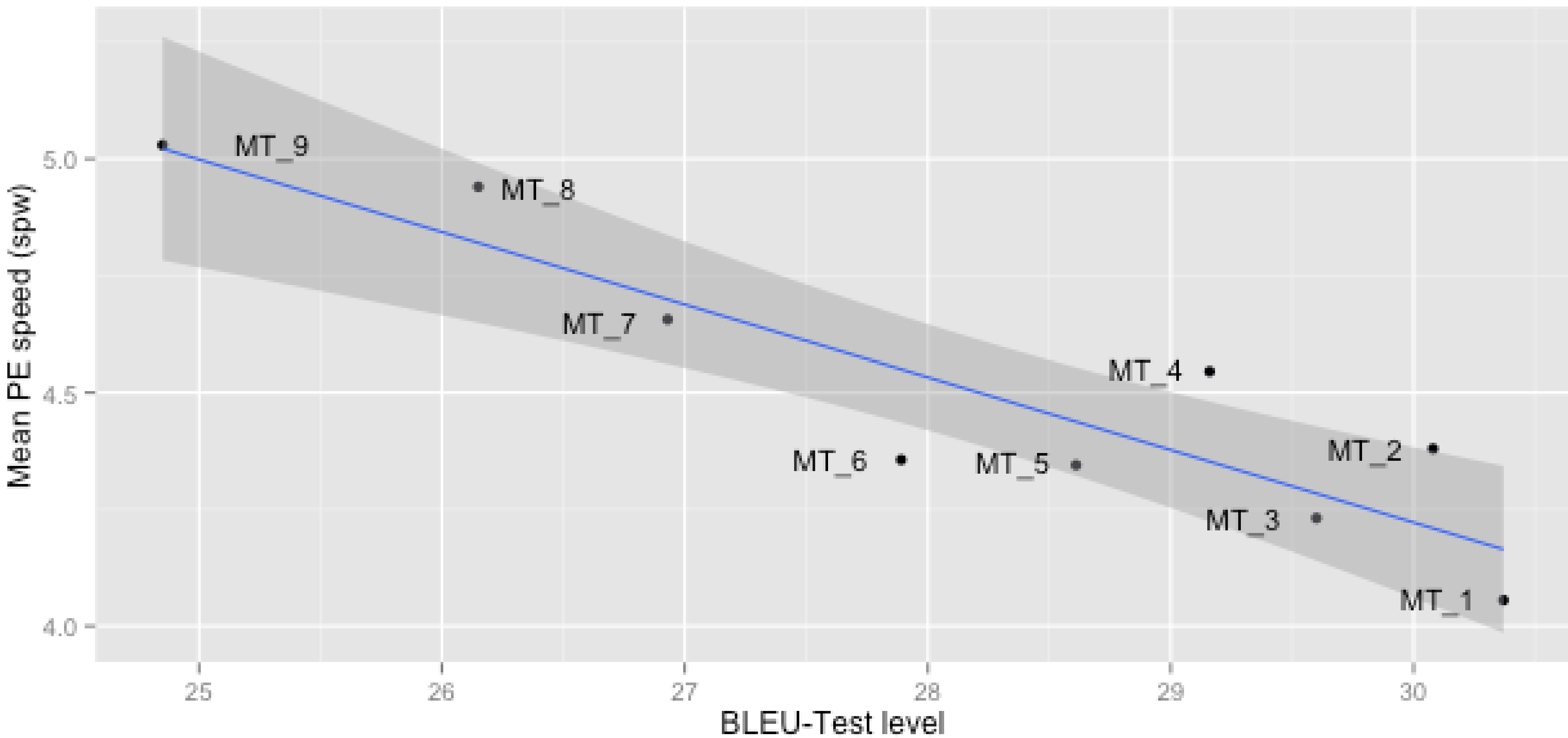
- Metrics inspired by this task
  - TER: based on number of editing steps
    Levenshtein operations (insertion, deletion, substitution) plus movement
  - HTER: manually construct reference translation for output, apply TER
    (very time consuming, used in DARPA GALE program 2005-2011)

Post-Editing Machine Translation



(source: Translated)

MT Quality and Productivity



BLEU against PE speed and regression line with 95% confidence bounds

+1 BLEU ↔ decrease in PE time of ~0.16 sec/word, or 3-4% speed-up

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

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
    $\rightarrow$ did person A **understand** the translation correctly?

statistical significance

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

- P-Levels
  - related to probability that there is a true difference
  - p-level $p < 0.01$ = more than 99% chance that difference is real
  - typcically used: p-level 0.05 or 0.01

- Confidence Intervals
  - given that the measured score is x
  - what is the true score (on a infinite size test set)?
  - interval $[x - d, x + d]$ contains true score with, e.g., 95% probability

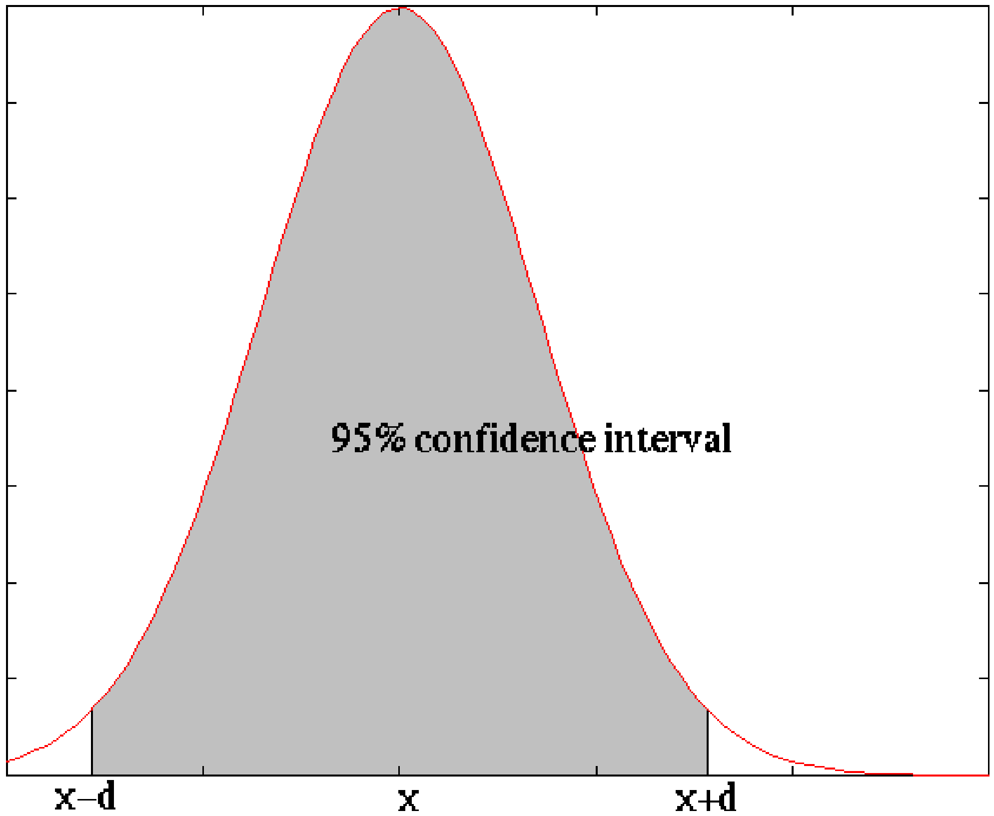
Computing Confidence Intervals

- Example
  - 100 sentence translations evaluated
  - 30 found to be correct

- True translation score?

  (i.e. probability that any randomly chosen sentence is correctly translated)

Normal Distribution



true score lies in interval $[\bar{x} - d, \bar{x} + d]$ around sample score $\bar{x}$
with probability 0.95

Confidence Interval for Normal Distribution

- Compute mean $\bar{x}$ and variance $\bar{s}^2$ from data

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

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

- True mean μ?

Student's t-distribution

- Confidence interval $p(\mu \in [\bar{x} - d, \bar{x} + d]) \geq 0.95$ computed by

$$d = t \, \frac{s}{\sqrt{n}}$$

- Values for t depend on test sample size and significance level:

Significance Level	Test Sample Size			
	100	300	600	∞
99%	2.6259	2.5923	2.5841	2.5759
95%	1.9849	1.9679	1.9639	1.9600
90%	1.6602	1.6499	1.6474	1.6449

Example

- Given
  - 100 sentence translations evaluated
  - 30 found to be correct

- Sample statistics
  - sample mean $\bar{x} = \frac{30}{100} = 0.3$
  - sample variance $s^2 = \frac{1}{99}(70 \times (0 - 0.3)^2 + 30 \times (1 - 0.3)^2) = 0.2121$

- Consulting table for t with 95% significance $\rightarrow$ 1.9849

- Computing interval $d = 1.9849 \frac{0.2121}{\sqrt{100}} = 0.042 \rightarrow [0.258; 0.342]$

Pairwise Comparison

- Typically, absolute score less interesting
- More important
  - Is system A better than system B?
  - Is change to my system an improvement?
- Example
  - Given a test set of 100 sentences
  - System A better on 60 sentence
  - System B better on 40 sentences
- Is system A really better?

Sign Test

- Using binomial distribution
  - system A better with probability p_A
  - system B better with probability p_B $(= 1 - p_A)$
  - probability of system A better on k sentences out of a sample of n sentences

$$\binom{n}{k} p_A^k \, p_B^{n-k} = \frac{n!}{k!(n-k)!} p_A^k \, p_B^{n-k}$$

- Null hypothesis: $p_A = p_B = 0.5$

$$\binom{n}{k} p^k \, (1-p)^{n-k} = \binom{n}{k} 0.5^n = \frac{n!}{k!(n-k)!} 0.5^n$$

Examples

n	$p \leq 0.01$		$p \leq 0.05$		$p \leq 0.10$	
5	-	-	-	-	$k = 5$	$\frac{k}{n} = 1.00$
10	$k = 10$	$\frac{k}{n} = 1.00$	$k \geq 9$	$\frac{k}{n} \geq 0.90$	$k \geq 9$	$\frac{k}{n} \geq 0.90$
20	$k \geq 17$	$\frac{k}{n} \geq 0.85$	$k \geq 15$	$\frac{k}{n} \geq 0.75$	$k \geq 15$	$\frac{k}{n} \geq 0.75$
50	$k \geq 35$	$\frac{k}{n} \geq 0.70$	$k \geq 33$	$\frac{k}{n} \geq 0.66$	$k \geq 32$	$\frac{k}{n} \geq 0.64$
100	$k \geq 64$	$\frac{k}{n} \geq 0.64$	$k \geq 61$	$\frac{k}{n} \geq 0.61$	$k \geq 59$	$\frac{k}{n} \geq 0.59$

Given n sentences
system has to be better in at least k sentences
to achieve statistical significance at specified p-level

Bootstrap Resampling

- Described methods require score at sentence level
- But: common metrics such as BLEU are computed for whole corpus
- Sampling
  1. test set of 2000 sentences, sampled from large collection
  2. compute the BLEU score for this set
  3. repeat step 1–2 for 1000 times
  4. ignore 25 highest and 25 lowest obtained BLEU scores

  $\rightarrow$ 95% confidence interval
- Bootstrap resampling: sample from the same 2000 sentence, with replacement