Machine Translation Evaluation

Sara Stymne

2017-03-29

Partly based on Philipp Koehn's slides for chapter 8

Why Evaluation?

- How good is a given machine translation system?
- Which one is the best system for our purpose?
- How much did we improve our system?
- How can we tune our system to become better?
- Hard problem, since many different translations acceptable
  $\rightarrow$ semantic equivalence / similarity

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)

Which translation is best?

Source	Färjetransporterna har minskat med 20,3 procent i år.
Gloss	The-ferry-transports have decreased by 20.3 percent in year.
Ref	Ferry transports are down by 20.3% in 2008.

Which translation is best?

Source	Färjetransporterna har minskat med 20,3 procent i år.
Gloss	The-ferry-transports have decreased by 20.3 percent in year.
Ref	Ferry transports are down by 20.3% in 2008.
Sys1	The ferry transports has reduced by 20.3% in year.
Sys2	This year, the reduction of transports by ferry is 20,3 procent.
Sys3	Färjetransporterna are down by 20.3% this year.
Sys4	Ferry transports have a reduction of 20.3 percent in year.
Sys5	Transports are down by 20.3%.

Evaluation Methods

- Subjective judgments by human evaluators
- Task-based evaluation, e.g.:
  - How much post-editing effort?
  - Does information come across?
- Automatic evaluation metrics
- Quality estimation

Human vs Automatic Evaluation

- Human evaluation is
  - Ultimately what we are interested in, but
  - Very time consuming
  - Not re-usable
  - Subjective
- Automatic evaluation is
  - Cheap and re-usable, but
  - Not necessarily reliable

Human evaluation

- Adequacy/Fluency (1 to 5 scale)
- Ranking of systems (best to worst)
- Yes/no assessments (acceptable translation?)
- SSER – subjective sentence error rate ("perfect" to "absolutely wrong")
- Usability (Good, useful, useless)
- Human post-editing time
- Error analysis

Adequacy and Fluency

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

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 target language?
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

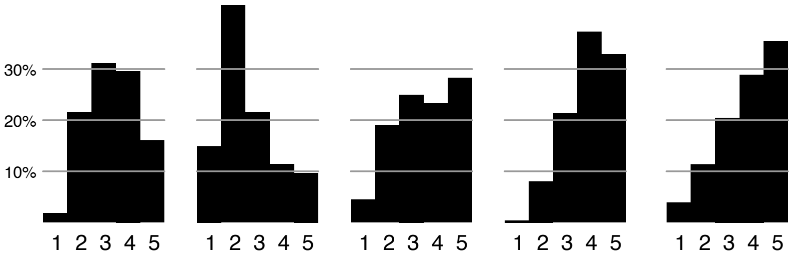
Fluency	
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Judge adequacy and fluency!

Source	Färjetransporterna har minskat med 20,3 procent i år.
Gloss	The-ferry-transports have decreased by 20.3 percent in year.
Ref	Ferry transports are down by 20.3% in 2008.
Sys4	Ferry transports have a reduction of 20.3 percent in year.
Sys5	Transports are down by 20.3%.
Sys6	This year, of transports by ferry reduction is percent 20.3.

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
- 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:

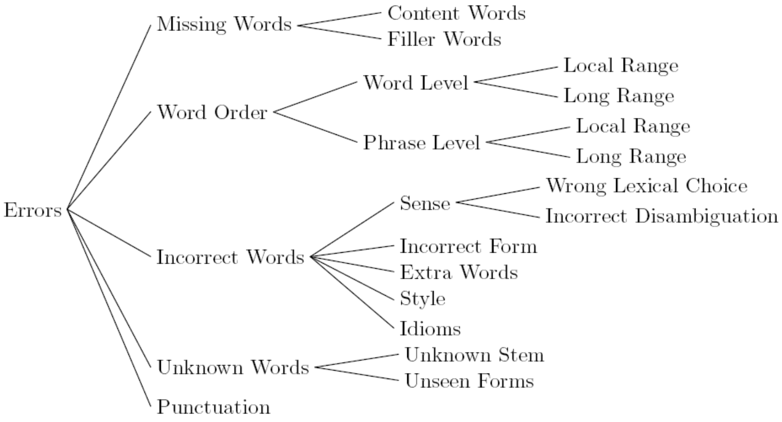
Evaluation type	$P(A)$	$P(E)$	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226
Sentence ranking	.582	.333	.373

Error Analysis

- Analysis and classification of the errors from an MT system
- Many general frameworks for classification exists, e.g.
  - Flanagan, 1994
  - Vilar et al. 2006
  - Costa-jussà et al. 2012
- It is also possible to analyse specific phenomena, like compound translation, agreement, pronoun translation, ...

Example Error Typology

Vilar et al.



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
- Some issues:
  - time consuming
  - depends on skills of translator/post-editor

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

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 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

Metrics – overview

- Precision-based
  - BLEU, NIST, ...
- F-score-based
  - Meteor, ...
- Error rates
  - WER, TER, PER, ...
- Using syntax/semantics
  - PosBleu, Meant, DepRef, ...
- Using machine learning
  - SVM-based techniques, TerrorCat

Metrics – overview

- Precision-based
  - **BLEU, NIST**, …
- F-score-based
  - **Meteor**, …
- Error rates
  - **WER, TER**, PER, …
- Using syntax/semantics
  - PosBleu, Meant, DepRef, …
- Using machine learning
  - SVM-based techniques, TerrorCat

Precision and Recall of Words

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

REFERENCE: Israeli officials are responsible for airport security

- Precision $$\frac{correct}{output\text{-}length} = \frac{3}{6} = 50\%$$
- Recall $$\frac{correct}{reference\text{-}length} = \frac{3}{7} = 43\%$$
- F-measure $$\frac{precision \times recall}{(precision + recall)/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

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%	86%
f-measure	46%	92%

flaw: no penalty for reordering

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)

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

- 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 (usually)
- Example

SYSTEM: Israeli officials | responsibility of | 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

NIST

- Similar to Bleu in that it measures N-gram precision
- Differences:
  - Arithmetic mean (not geometric)
  - Less frequent n-grams are weighted more heavily
  - Different brevity penalty
  - $N = 5$

METEOR: Flexible Matching

- Partial credit for matching stems

SYSTEM	Jim walk home
REFERENCE	Joe walks home

- Partial credit for matching synonyms

SYSTEM	Jim strolls home
REFERENCE	Joe walks home

- Use of paraphrases
- Different weights for content and function words (later versions)

METEOR

- Both recall and precision
- Only unigrams (not higher n-grams)
- Flexible matching (Weighted P and R)
- Fluency captured by a penalty for high number of chunks

$$F_{mean} = \frac{PR}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

$$Penalty = 0.5 * \gamma \cdot \left(\frac{\#chunks}{\#unigrams\_matched} \right)^{\beta}$$

$$Meteor = (1 - Penalty) \cdot F_{mean}$$

METEOR: tuning

- Meteor parameters can be tuned based on human judgments

Language	α	β	γ	δ	w_{exact}	w_{stem}	w_{syn}	w_{par}
Universal	.70	1.40	.30	.70	1.00	–	–	.60
English	.85	.20	.60	.75	1.00	.60	.80	.60
French	.90	1.40	.60	.65	1.00	.20	–	.40
German	.95	1.00	.55	.55	1.00	.80	–	.20

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

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

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%

Other error rates

- PER – position-independent word error rate
  - Does not consider the order of words
- TER – translation edit rate
  - Adds the operation SHIFT – the movement of a contigous sequence of words an arbritray distance
- SER – sentence error rate
  - The percentage of sentences that are identical to reference sentences

Metrics using syntax/semantics

- Posbleu, Bleu calculated on part-of-speech
- ULC – Overlap of:
  - shallow parsing
  - dependency and consituent parsing
  - named entities
  - semantic roles
  - discourse representation structures
- Using dependency structures
- Meant, semantic roles
- Considerations:
  - parsers/taggers do not perform well on misformed MT output
  - parsers/tagger not available for all languages

Critique of Automatic Metrics

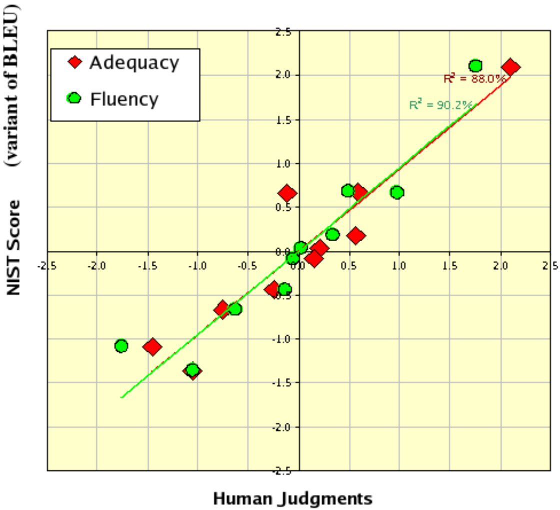
- Ignore relevance of words
  (names and core concepts more important than determiners and punctuation)
- Operate on local level
  (do not consider overall grammaticality of the sentence or sentence meaning)
- Scores are meaningless
  (scores very test-set specific, absolute value not informative)
- 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?

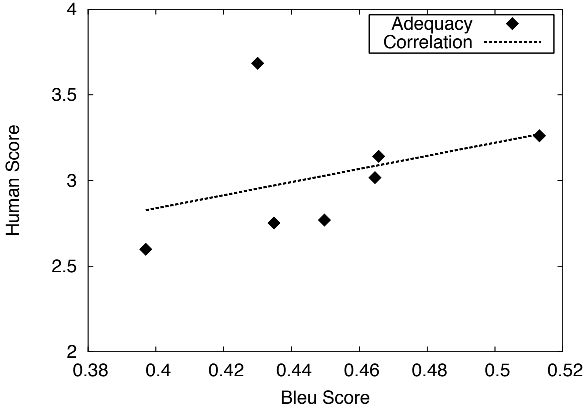
→ Yes, if they correlate with human judgement

Correlation with Human Judgement



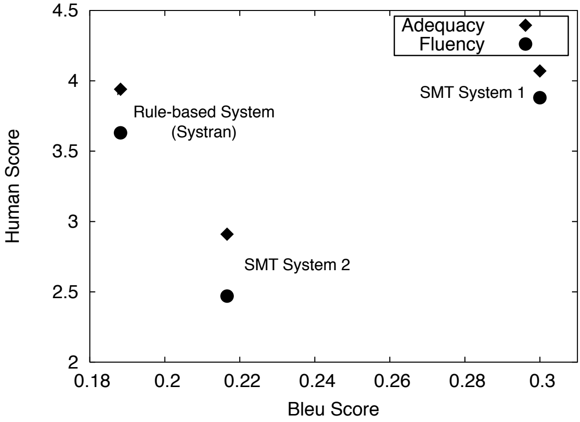
Evidence of Shortcomings of Automatic Metrics

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)

Correlations of metrics with human ranking

Metric	de-en	en-de
BLEU	.90	.79
METEOR	.96	.88
TER	.83	.85
WER	.67	.83
TERRORCAT	.96	.95
DEPREF-ALIGN	.97	–

(System level, WMT 2013)

Correlations of metrics with human ranking

Metric	de-en	en-de
BLEU	.23	.18
METEOR	.26	.24
TERRORCAT	.25	.21
DEPREF-ALIGN	.26	–

(Segment level, WMT 2013)

Automatic Metrics: Conclusions

- Automatic metrics essential tool for system development
- Not fully suited to rank systems of different types
- Reasonable results on system level evaluation, but not on sentence level
- Evaluation metrics still open challenge

Quality Estimation

- For standard automatic metrics, a reference translation is needed
- In a translation scenario, we do not have reference translations
- It is very useful for a translator who is presented MT output to know:
  - Is it good enough as it is
  - Can it be easily edited
  - Can it be edited with some effort
  - Is it completely useless
- This task is called quality estimation

Quality Estimation – Details

- Automatic evaluation without a reference
- Typically modelled as a machine learning task
- Using features such as:
  - How long is the sentence?
  - What is the length difference between the source and target?
  - How common are the words and n-grams in the source sentence?
  - How ambiguous are the words in the source sentence?
  - How many punctuation marks are there in the sentence?
- Train on judgments of fluency/adequacy, post-editing effort, or post-editing time

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 an infinite size test set)?
  - interval $[x - d, x + d]$ contains true score with, e.g., 95% probability

Pairwise Comparison

- Typically, we want to know if one system is better than another
  - 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	$p \leq 0.01$		$p \leq 0.05$	
5	-	-	-	-
10	$k = 10$	$\frac{k}{n} = 1.00$	$k \geq 9$	$\frac{k}{n} \geq 0.90$
20	$k \geq 17$	$\frac{k}{n} \geq 0.85$	$k \geq 15$	$\frac{k}{n} \geq 0.75$
50	$k \geq 35$	$\frac{k}{n} \geq 0.70$	$k \geq 33$	$\frac{k}{n} \geq 0.66$
100	$k \geq 64$	$\frac{k}{n} \geq 0.64$	$k \geq 61$	$\frac{k}{n} \geq 0.61$

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

Data-driven Significance Testing

- Described methods require score at sentence level
- But: common metrics such as BLEU are computed for whole corpus
- Data-driven methods are typically used
- Bootstrap resampling
  - Sample sentences from the test set, with replacement
- Approximate randomization
  - Scramble sentences between the two systems that you compare

Summary

- MT evaluation is hard
- Human evaluation is expensive
- Automatic evaluation is cheap, but not always fair
- What is typically used in MT research:
  - Bleu!
  - Maybe another/several other metrics (typically Meteor, TER)
  - Maybe some human judgments
    - Ranking of systems
    - Targeted analysis of specific phenomenon
- $\rightarrow$ Be careful when you argue about MT quality!

Outlook

- Next week: MT in practice
  - Guest lecture, Convertus (Commercial MT solutions in Uppsala)
  - Lab 1: Evaluation (Written lab report)
    REMEMBER: sign up for lab pairs!
- Coming weeks:
  - Introduction to SMT
  - Lab 2: Word-based models
    - 1st part: oral examination in class
      be present and active for the full session
      (write report if absent/passive)
    - 2nd part: written lab report