INF5820/INF9820

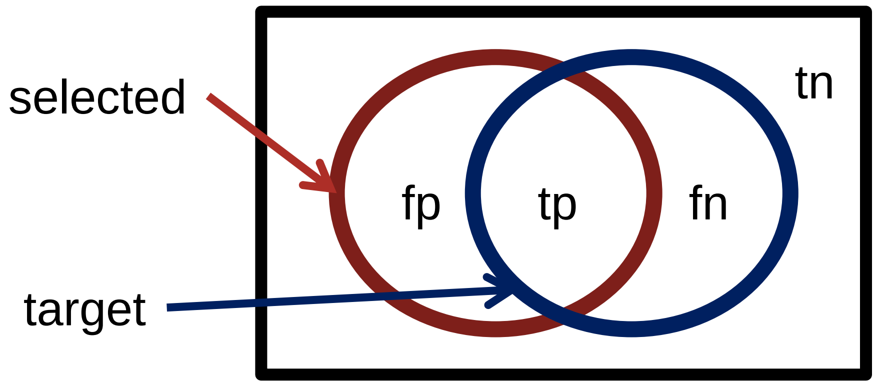
LANGUAGE TECHNOLOGICAL APPLICATIONS

Jan Tore Lønning, Lecture 3, 5 Sep.
jtl@ifi.uio.no

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. Word precision and recall (from last week)
   2. BLEU
   3. Alternatives
   4. Evaluation evaluation
   5. Criticism
2. Evaluation of applied MT-systems

Common evaluation measures in LT

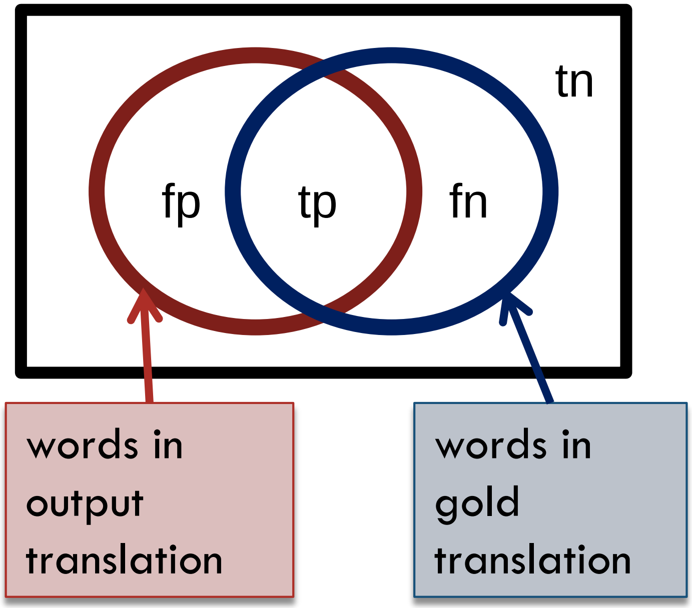


		Actual (gold)	
		target	Not target
System perform	selected	tp: True positive	fp: False positive
	Not selected	fn: False negative	tn: True negative

- Recall $= \dfrac{tp}{tp + fn}$
- Precision $= \dfrac{tp}{tp + fp}$
- F-score $= \dfrac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$
- $F_1 = \dfrac{1}{0.5\frac{1}{P} + (1-0.5)\frac{1}{R}} = \dfrac{2PR}{R+P}$

Adapting P, R, F to MT-eval

- Precision $= \dfrac{correct}{output.length}$
- Recall $= \dfrac{correct}{ref.length}$
- $F_1 =$



$$\frac{2}{\frac{1}{R}+\frac{1}{P}} = \frac{2}{\frac{ref.length}{correct}+\frac{output.length}{correct}} = \frac{2correct}{output.length + ref.length}$$

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%~~ $\frac{6}{7} \approx 0.86$
f-measure	46%	~~100%~~ $\frac{12}{13} \approx 0.92$

flaw: no penalty for reordering

Position-independent error rate

- Similar measure to (word) recall+precision
- Reports mistakes – not correctness
- We skip the details - formula

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

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

Levenshtein distance used in
- spell-checking
- OCR
- Translation memory

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%

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. Word precision and recall (from last week)
   2. BLEU
   3. Alternatives
   4. Evaluation evaluation
   5. Criticism
2. Evaluation of applied MT-systems

BLEU

- A Bilingual Evaluation Understudy Score
- Main ideas:
  - Use several reference translations
  - Count precision of n-grams:
    - For each n-gram in output:
      does it occur in at least one reference?
  - Don’t count recall but use a penalty for brevity
    - Why not recall?

BLEU

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram, C, C.refs)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram, C)}$$

- Candidates:
  - the set of sentences output by trans. system
- Count(n-gram, C):
  - the number of times *n-gram* occurs in C
- $Count_{clip}$(n-gram, C, C.refs):
  - the number of times the *n.gram* occurs in both
    - C and
    - the reference translation for the same sentence
    - where *n.gram* occurs most frequent

- Technicality:
  - If the same n-gram has several occurrences in a candidate translation sentence, it should not be counted more times than the number of occurrences in the reference sentence with the largest number of occurrences of the same n-gram.

Example, p_3

- Hyp, C:
  - One of the girls gave one of the boys one of the boys.
- C-Refs:
  - A girl gave a boy one of the toy cars
  - One of the girls gave a boy one of the cars.
- Count_clip(one of the, C, C-refs)=2

one of the	of the girls	the girls gave	girls gave one
2 (3)	1	1	1

gave one of	of the boys	the boys one	boys one of
0 (1)	0 (2)	0 (1)	0 (1)

- $P_3 = 5/11$

BLEU

- How to combine the n-gram precisions?

$$p_1 \times p_2 \times \cdots \times p_n = \prod_{i=1}^{n} p_i$$

- Remember

$$\ln(\prod_{i=1}^{n} p_i) = \ln(p_1 \times p_2 \times \cdots \times p_n) = \ln(p_1) + \ln(p_2) + \cdots + \ln(p_n) = \sum_{i=1}^{n} \ln p_i$$

- One can add weights, typically *ai* = 1/*n*

$$\ln(p_1^{a1} \times p_2^{a2} \times \cdots \times p_n^{an}) = a1 \ln(p_1) + a2 \ln(p_2) + \cdots + an \ln(p_n)$$

- How long n-grams?
  - Max 4-grams seems to work best

Brevity penalty

- c is the length of the candidates
- r is the length of the reference translations:
  - for each C choose the R most similar in length

$$c = \sum_{C \in Candidates} length(C)$$

$$r = \sum_{C \in Candidates} length(R.sim.C)$$

- Penalty applies if c < r:
  - BP = 1 if $c \geq r$
  - BP = $e^{(1-r/c)}$ otherwise

This is correct
Error in K:SMT

- $$BLEU = BP \cdot \exp \sum_{i=1}^{n} w_n \ln p_i$$
- $$\ln BLEU = \min(1-\frac{r}{c},0) + \sum_{i=1}^{n} w_n \ln p_i$$

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. Word precision and recall (from last week)
   2. BLEU
   3. Alternatives
   4. Evaluation evaluation
   5. Criticism
2. Evaluation of applied MT-systems

NIST score

- National Institute of Standards and Technology
- Evaluated BLEU score further
- Proposed an alternative formula:
  - N-grams are weighed by their inverse frequency
  - Sums (instead of products) of counts over n-grams
  - Modified Brevity Penalty
- Freely available software

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. Word precision and recall (from last week)
   2. BLEU
   3. Alternatives
   4. Evaluation evaluation
   5. Criticism
2. Evaluation of applied MT-systems

Evaluating the automatic evaluation

- Is the automatic evaluation correct?
- Yes, if it gives the same results as human translators.
  - Same results best measured as ranking of MT systems

BLEU – original paper

Figure 2: Machine and Human Translations

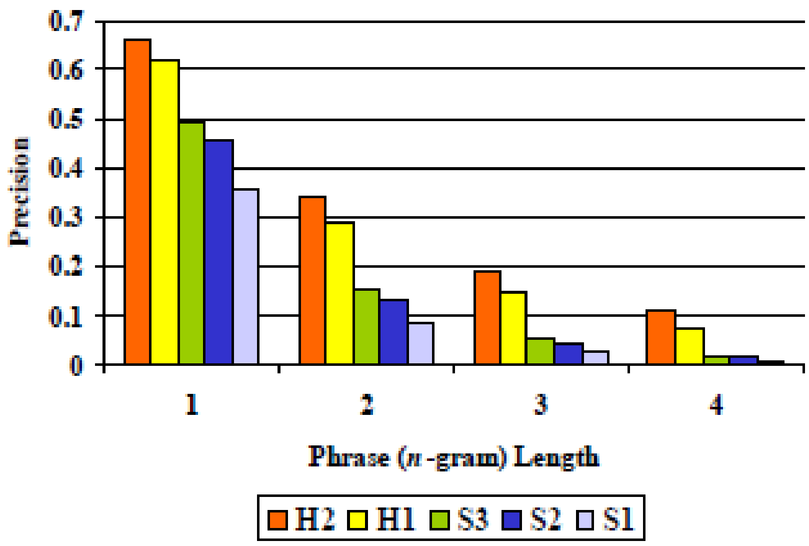
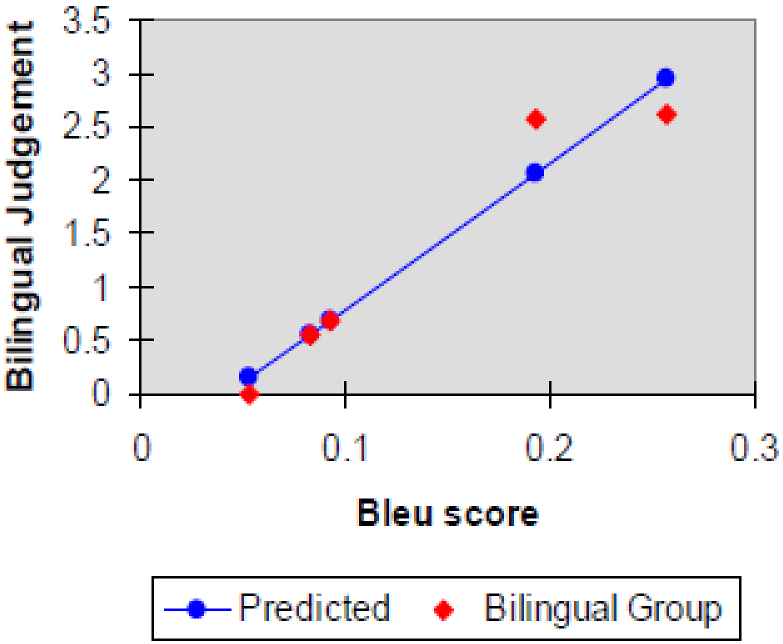


Figure 6: BLEU predicts Bilingual Judgments



H1, H2 – 2 different human translations
S1, S2, S3 – different MT systems

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:

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

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

Machine Translation Evaluation 2

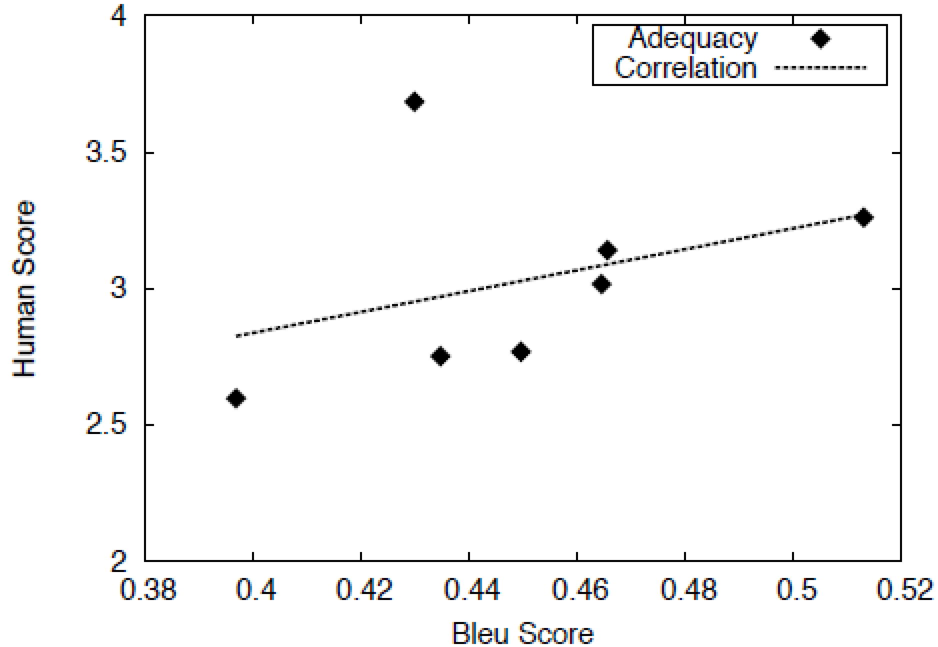
1. Automatic MT-evaluation:
    1. Word precision and recall (from last week)
    2. BLEU
    3. Alternatives
    4. Evaluation evaluation
    5. Criticism
2. Evaluation of applied MT-systems

Shortcomings of automatic MT

- Re-evaluating the Role of BLEU in Machine Translation Research, 2006
  - Chris Callison-Burch, Miles Osborne, Philipp Koehn
- Theoretically:
  - From a reference translation one may
  - Construct a string of words, which:
  - Gets a high BLEU score
  - Is gibberish
- Empirically: (next slides)

Evidence of Shortcomings of Automatic Metrics

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



Automatic evaluation

- ☺ Cheap
- ☺ Reusable in development phase
- ☺ A touch of objectivity
- ☺ Useful tool for machine learning, e.g. reranking

- ☹ Does not measure MT quality,
  only (more or less) correlated with MT quality
- ☹ Favors statistical approaches, disfavors humans
- ☹ The numbers don't say anything across different evaluations
  - ☹ Depends on number and type of reference translations
- ☹ Danger of system tuning towards BLEU on the cost of quality
  - ☹ In particular in machine learning

Hypothesis testing

- You may skip sec. 8.3
- Though:
  - 8.3.1 for they who have INF5830
  - 8.3.2, when you have 2 different systems
    - You might evaluate first one system, then the other on the whole material and compare the results
    - Often better: Compare item by item which system is the better and do statistics on the results

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. Word precision and recall (from last week)
   2. BLEU
   3. Alternatives
   4. Evaluation evaluation
   5. Criticism
2. Evaluation of applied MT-systems

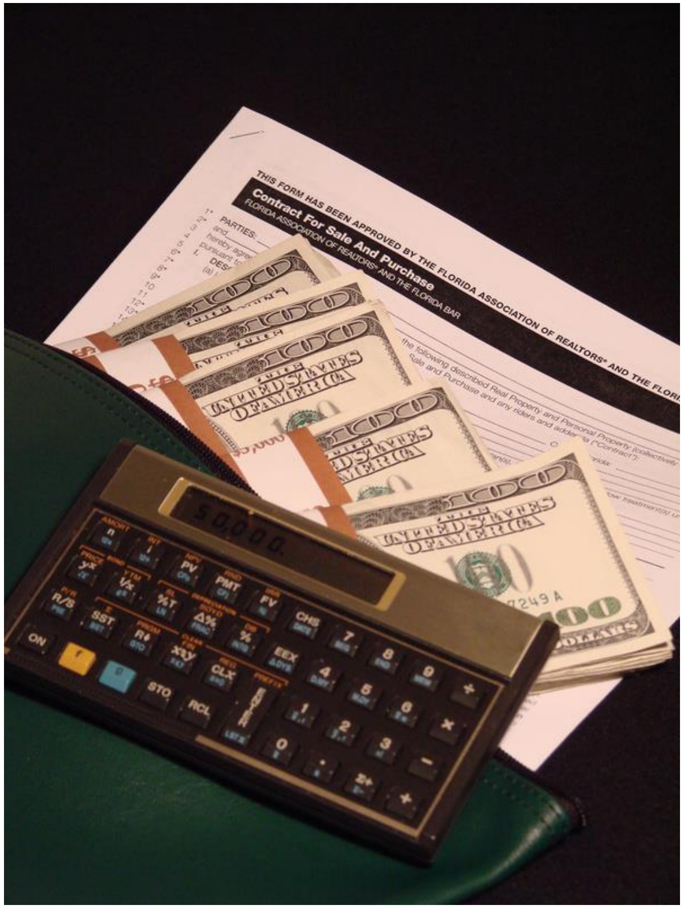
MT Evaluation – a broader perspective

- (Human) MT-evaluation:
  - Long history,
    - e.g. the ALPAC-report 1966
  - Research field on its own

- Evaluation distinctions:
  - A larger system with MT as a part vs the MT module
  - The whole MT system or its parts
    - ”black box” vs ”glass box”
  - Text vs task (instructions for assembling a bookcase)
  - Text vs reading understanding

MT Evaluation from outside

- What are we willing to give up (no FAHQT?)
- The consumer perspective:
  - Price
  - Speed
  - Covered language pairs
  - Maintenance cost
  - Cost and speed of post-editing
  - Training costs



Conclusions

- Evaluation of MT can be done with respect to various properties
- Particularly quality
- Automatic evaluation
  - Pros
  - Cons