INF5820/INF9820

LANGUAGE TECHNOLOGICAL APPLICATIONS

Jan Tore Lønning, Lecture 3, 7 Sep., 2016
jtl@ifi.uio.no

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. BLEU
   2. Alternatives
   3. Evaluation evaluation
   4. Criticism
2. Starting STMT
   1. The noisy channel model
   2. Language models (n-grams)

Last week

- Human evaluation
- Machine evaluation
  - Recall and precision
  - Word error rate
  - BLEU

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

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_1 and p_2

- 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', C, C-refs)=2

one	of	the	girls	gave	boys					
2 (3)	2 (3)	2 (3)	1	1	0(2)					

- $P_1 = 8/13$

one of	of the	the girls	girls gave	gave one	the boys	boys one
2 (3)	2 (3)	1	1	0 (1)	0(2)	0 (1)

- $P_2 = 6/12$

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	0 (1)

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

- $P_3 = 4/11$

Example continued

$$\prod_{i=1}^{4} p_i = p_1 \cdot p_2 \cdot p_3 \cdot p_4 = \frac{8}{13} \cdot \frac{6}{12} \cdot \frac{4}{11} \cdot \frac{2}{10} \approx 0.02238$$

$$\left(\prod_{i=1}^{4} p_i\right)^{\frac{1}{4}} \approx 0.02238^{\frac{1}{4}} \approx 0.39$$

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

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

Use logarithms to avoid underflow!

BLEU-4

$$\text{BLEU}{-}4 = exp\left(\min\left(1 - \frac{r}{c}, 0\right) \sum_{i=1}^{4} \frac{1}{4} \ln p_i \right)$$

$$\text{BLEU}{-}4 = \min\left(e^{\left(1-\frac{r}{c}\right)}, 1\right) \left(\prod_{i=1}^{4} p_i \right)^{\frac{1}{4}}$$

Machine Translation Evaluation 2

1. Automatic MT-evaluation:
    1. BLEU
    2. Alternatives
    3. Evaluation evaluation
    4. Criticism
2. Starting STMT
    1. The noisy channel model
    2. Language models (n-grams)

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

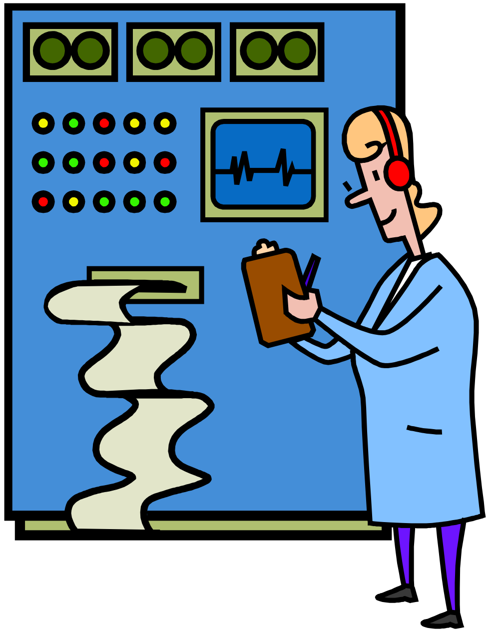
Machine Translation Evaluation 2

1. Automatic MT-evaluation:
   1. BLEU
   2. Alternatives
   3. Evaluation evaluation
   4. Criticism
2. Starting STMT
   1. The noisy channel model
   2. Language models (n-grams)

Evaluating the automatic evaluation

- Is the automatic evaluation correct?
- Yes, if it gives the same results as human evaluators.
  - Best measured as ranking of MT systems:

Does BLEU rank a set of MT systems in the same order as human evaluators?



BLEU – original paper

Figure 2: Machine and Human Translations

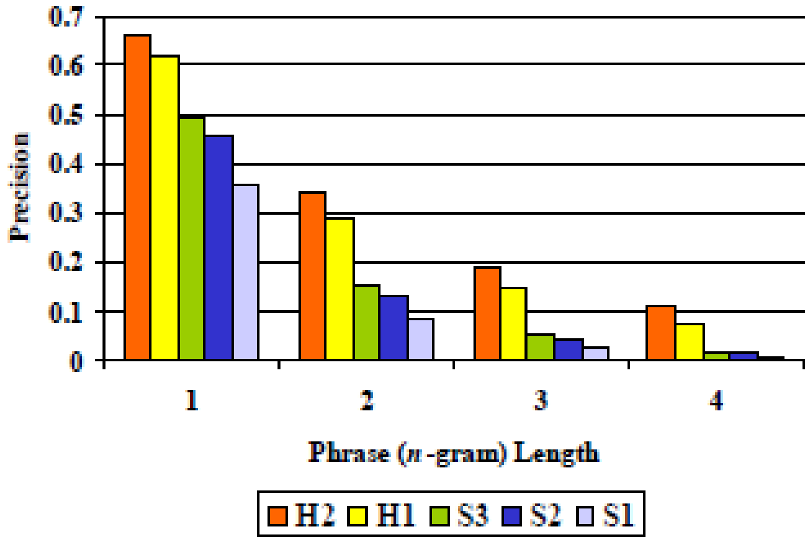
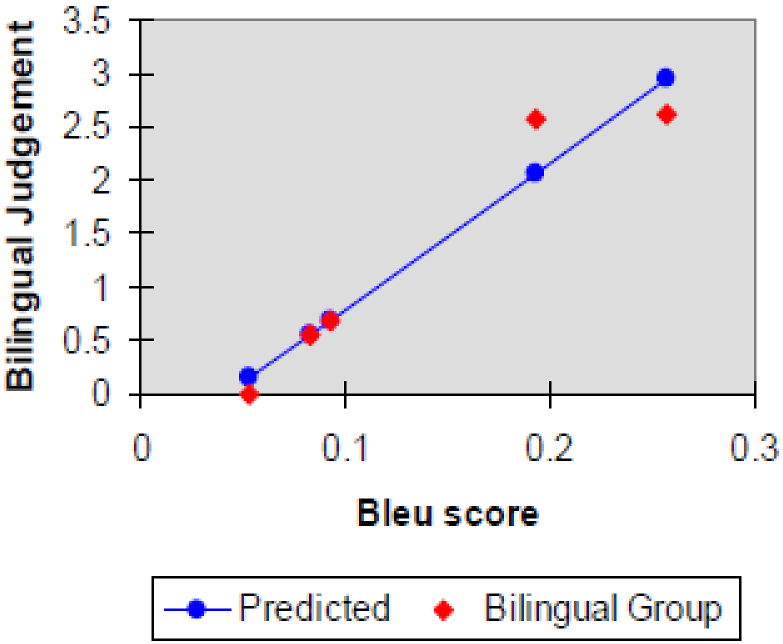


Figure 6: BLEU predicts Bilingual Judgments



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:

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

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

Machine Translation Evaluation 2

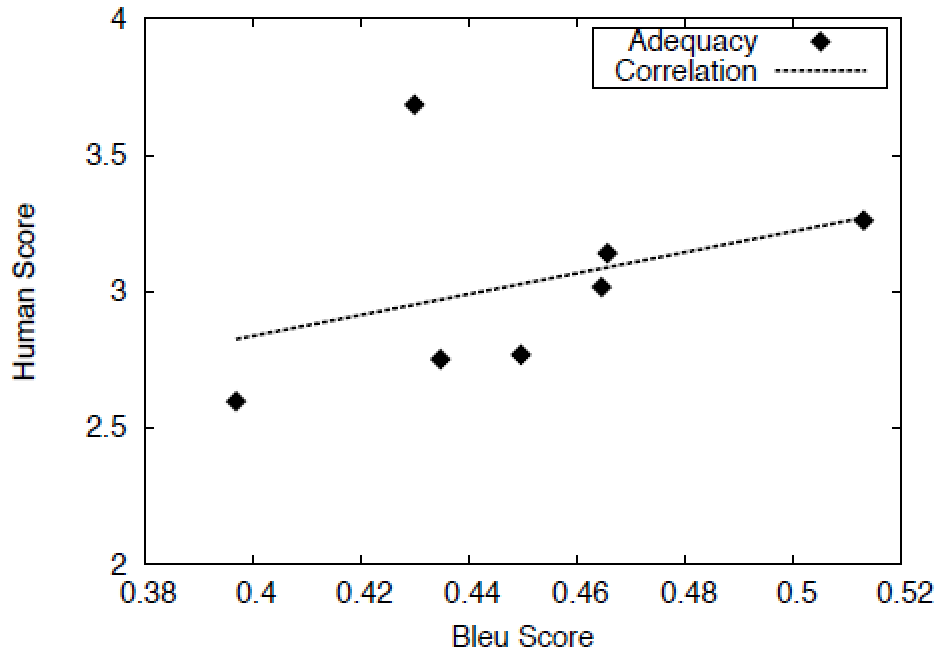
1. Automatic MT-evaluation:
   1. BLEU
   2. Alternatives
   3. Evaluation evaluation
   4. Criticism
2. Starting STMT
   1. The noisy channel model
   2. Language models (n-grams)

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. BLEU
    2. Alternatives
    3. Evaluation evaluation
    4. Criticism
2. Starting STMT
    1. The noisy channel model
    2. Language models (n-grams)

SMT example

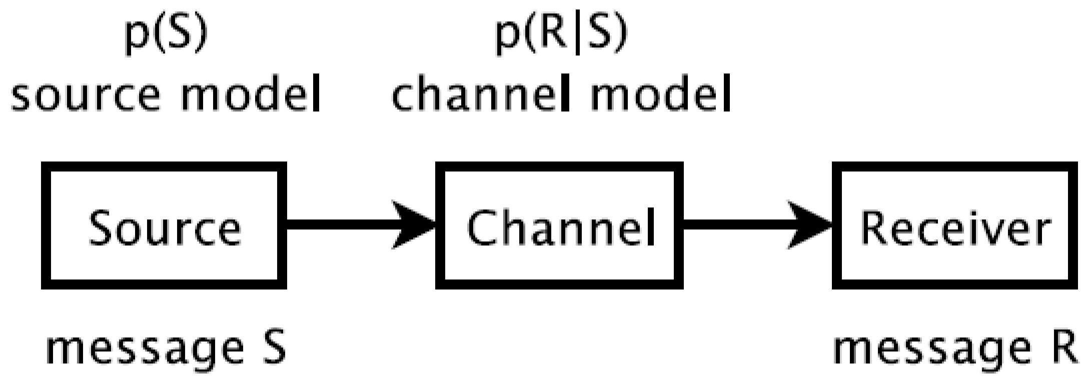
En	kokk	lagde	en	rett	med	bygg	.
a 0.9	chef 0.6	made 0.3	a 0.9	right 0.19	with 0.4	building 0.45	
…	cook 0.3	created 0.25	…	straight 0.17	by 0.3	construction 0.33	
	…	prepared 0.15		court 0.12	of 0.2	barley 0.11	
		constructed 0.12		dish 0.11	…	…	
		cooked 0.05		course 0.07			
		…		…			

Similarly for:

- pos 0-2 (2x3)
- pos 1-3
- pos 2-4
- pos 3-5 (4x5)
- pos 6-8

Pos4 – pos 6 (1x3x3 many)		Pos5 – pos 7 (5x3x3 many)	
a right with	2.7×10^{-12}	right with building	1.7×10^{-18}
a right of	1.5×10^{-10}	right with construction	5.4×10^{-18}
a right by	9.7×10^{-12}	right with barley	8.7×10^{-19}
…		…	
a course of	1.5×10^{-14}	course of barley	1.5×10^{-16}

Noisy Channel Model



- Applying Bayes rule also called noisy channel model
  - we observe a distorted message R (here: a foreign string **f**)
  - we have a model on how the message is distorted (here: translation model)
  - we have a model on what messages are probably (here: language model)
  - we want to recover the original message S (here: an English string **e**)