Lecture 6.2: automatic evaluation of MT

Why automatic?

- Cost
- Speed
- Without “emotion” 😉
- How to evaluate the evaluation?
  - Does it correspond to human judgment?
  - If *system A* has a better score than *system B* humans should prefer its translations

Metrics

- BLEU: **Bi**Lingual **E**valuation **U**nderstudy
  - Most popular metric
  - Correlates with human judgment… sometimes!
  - Based on words (and n-grams) overlap between reference and MT output
    - Rewards words in the same order
    - “clipped word count” to ”punish” repeated words in MT
    - Brevity penalty to “punish” too short/too long MT

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation". ACL-2002: *40th Annual meeting of the Association for Computational Linguistics.* pp. 311–318.

BLEU METRIC

- $p_n = \dfrac{\sum_{c\in\{candidates\}} \sum_{ngram\in C} Count_{clip}(ngram)}{\sum_{C'\in\{candidates\}} \sum_{ngram'\in C'} Count\ \ (ngram')}$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

BP: brevity
c: total length of candidate translation corpus
r: test corpus' effective reference length

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

N=number of words in n-grams (consecutive words form n-grams up to length *N*)
Usually N=4

- $BLEU = BP \times \exp(\sum_n \log p_n)$

BLEU

- Score between 0 and 1 (0-100%):
  - 0-no common words
  - 1-MT is exactly one of the reference
- No scoring at sentence level (lack of 4 grams=>division by 0)
  - see smoothed BLEU
  - E.g. Moses Mert uses "sentence-bleu"
- No synonym, flexions or paraphrases
  - BLEUs(*)("this house has window", "these houses have windows") = 0.3
- No difference between words
  - BLEUs("this house has window", "this dog has legs") = 0.39
  - BLEUs("this house has window", "the house have window") = 0.39

BLEUs: sentence level BLEU

Other metrics

- METEOR
- Levenstein’s string edit distance
- Word error rate
- Translation Eror Rate

METEOR

- Uses stems and synonyms (paraphrases) for its similarity
- Pro: Usually more accurate than BLEU
- Cons: works only in some languages (e.g. paraphrases in es,cz,fr,en,de,ru)

RIBES

- Word rank-based metric that compares the ratio of contiguous and dis-contiguous word pairs between MT and reference
- Efficient metric for e.g. Japanese-English

Other metrics

- Levenstein’s string edit distance
- Word error rate
- Translation Edit Rate
- METEOR

Levenshtein edit distance

- The Levenshtein edit distance (in our context) between two sentences is the number of words we have to delete, insert or substitute to change one sentence into the other

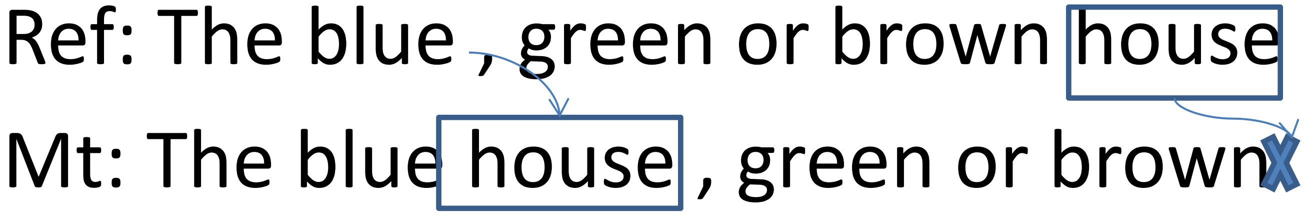
$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if} \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

- In automatic MT evaluation: a metric measuring the “distance” between reference and MT translation
- In post-edition evaluation: it represents the number of words that a post-editor changed to correct a MT
- This is a distance: between 0 and *n* (*n:* max number of words)

Word error rate (WER)

- WER is the Leveinshtein distance normalized by number of words.
- It is a ratio, between 0 (all words are the same) and 1 (all words are different)

F: La maison bleue , verte ou marron



Ref: The blue , green or brown house

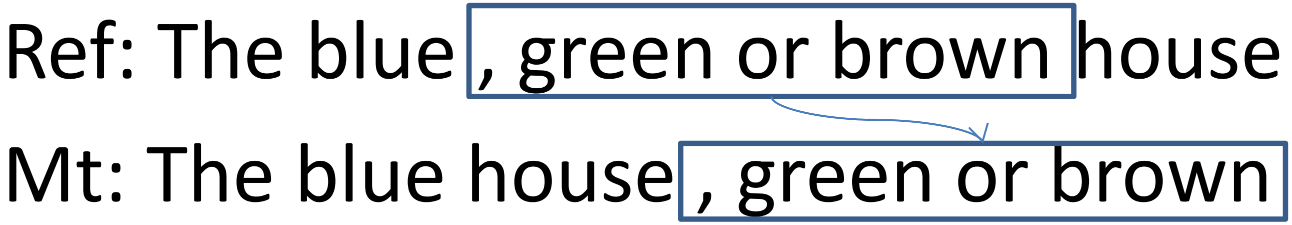
Mt: The blue house , green or brown ✗

- WER = (1 del + 1 ins) / 7 = 0.286

Translation error rate (TER)

- Same as WER but adds “shift” operation

F: La maison bleue , verte ou marron



Ref: The blue , green or brown house

Mt: The blue house , green or brown

- TER= 1 shift / 7 = 0.143