EVALUATION OF TRANSLATION QUALITY

Why do we want to evaluate translations?

- To compare systems
- To evaluate incremental changes and find fruitful avenues for improvement.

CRITERIA FOR EVALUATION

Humans evaluate MT quality on a five point scale

- **Fluency** of the target
  - 5 flawless
  - 4 good
  - 3 non-native
  - 2 disfluent
  - 1 incomprehensible

- **Adequacy**: how much of the information of the reference translation is kept in the target?
  - 5 All
  - 4 Most
  - 3 Much
  - 2 Little
  - 1 None

GOALS FOR AUTOMATIC EVALUATION

- Human evaluation is reliable, but costly and not reusable

Goal for automatic evaluation

- cost-free evaluation for incremental changes
- identify source of problem and analysis of errors
- measure correlates with human judgments (native speakers and professional translators)
- score is easy to interpret
- score allows comparison across systems

WHY IS IT HARD?

The general goal is to get as close as possible to human translations by comparing to reference translations.

- We cannot assume there is only one reference translation
- Exact match cannot be expected
- Translation is not transcriptions; so phrase can move around.

SOLUTION FOR AUTOMATIC EVALUATION

- Expect exact match to a <u>set</u> of reference translations
- Distance from reference translation is calculated on substrings (n-grams).
- Most commonly used measure is BLEU
    - multiple reference translations
    - n-gram exact match anywhere in the sentence
    - penalty for brevity

<u>Main idea</u>: the more n-grams shared with references, the better

BLEU SCORE

$$\text{BLEU} = \text{Brevity Penalty} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{Brevity Penalty} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

c is length of the corpus of reference translations

r length of "effective reference corpus"

Modified n-gram Precision

- p_n = precision for each n-gram length calculated by summing over the matches for every hypothesised translation in the corpus

- Each p_n is weighted by w_n. In practice all n-grams are equally weighted.

EXAMPLE

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Table 1: A set of four reference translations, and a hypothesis translation from the 2005 NIST MT Evaluation

1-grams: **American**, **Florida**, **Miami**, Orejuela, **appeared**, as, being, **calm**, carry, escorted, **he**, him, in, led, **plane**, quite, seemed, take, that, **the**, **to**, **to**, to, **was** , was, **which**, while, **will**, would, **,** **.**
2-grams: **American plane**, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, **, Florida**
3-grams: American plane that, American plane which, **Miami , Florida**, Miami in Florida, Orejuela appeared calm, Orejuela seemed quite, appeared calm as, appeared calm while, as he was, being escorted to, being led to, calm as he, calm while being, carry him to, escorted to the, he was being, he was led, him to Miami, in Florida ., led to the, plane that was, plane that would, plane which will, quite calm as, seemed quite calm, take him to, that was to, that would take, **the American plane**, the plane that, **to Miami ,**, to Miami in, to carry him, **to the American**, to the plane, was being led, was led to, was to carry, which will take, while being escorted, will take him, would take him, **, Florida .**

$$p_1 = \frac{15}{18} = .83$$

$$p_2 = \frac{10}{17} = .59$$

$$p_3 = \frac{5}{16} = .31$$

$$p_4 = \frac{3}{15} = .2$$

Table 2: The n-grams extracted from the refer-

EVALUATION OF EVALUATION SCORES

- Bleu is reliable because it has been shown to match human judgements in
  - ranking systems
  - distinguishing human from machine translations

- <u>Ranking consistency</u>: a good score must be consistent across its scores. It must be able to rank similar translations similarly.

  A measure of variability of BLUE scores according to the choice of documents and choice of reference translation shows good consistency.

BLEU CORRELATION WITH HUMAN EVALUATION

- Extensive evaluation on four source languages:
  - French, Japanese, Spanish, Chinese

⇓

English

  - 100 documents each
  - Human quality judgments are available for each of the documents

CORRELATION OF BLEU WITH HUMAN JUDGMENTS

CORPUS	SYSTEMS	ADEQUACY %	FLUENCY %
French	5 MT	95.7	99.7
Japanese	4 MT	97.8	85.6
Spanish	4 MT	97.5	97.2
Chinese	6 MT	95.2	97.1
	7 HT	70.5	16.6

Observations
- very high correlations with HJ
- lower correlation with HT because human translators are more varied
- lower correlation with Japanese because of lack of variation across systems

COMMENTARIES ON BLEU

- Experience shows that using multiple reference translations has very limited effect on quality of BLEU scores.

- BLEU calculated over longer segments (e.g. whole document) correlates equally well.

- Criticisms: improvement in BLEU scores is neither necessary nor sufficient to indicate improvement in translation.

  - BLEU imposes no constraints on which n-grams are matched to allow for change in word order and lexical choice
  - Completely wrong word orders can receive a good score
  - Different lexical choices and structures (as preferred by professional translators) are penalized.