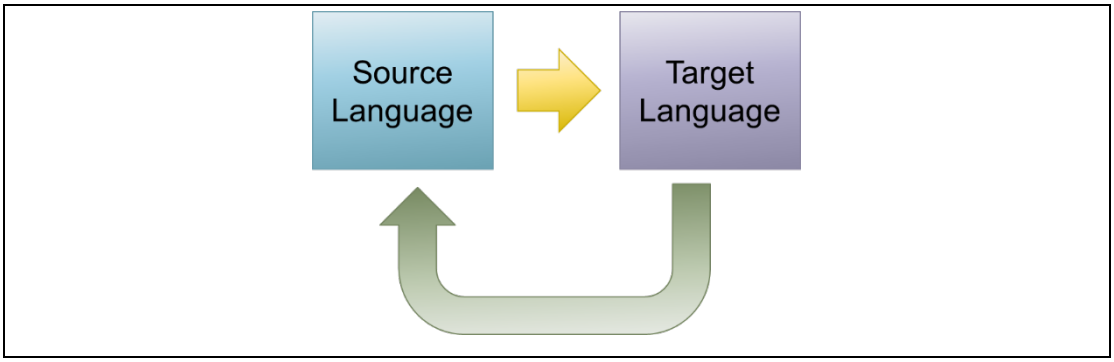
TRA6001 Translation Technology: Knowledge and Skills

Supplementary Handout 22

Evaluation of Machine Translation

A. Round-trip Translation

The source text is first translated into the target language by using MT, and the MT output is translated back into the source language. If the new source text is like the original text, it is likely that the MT system concerned is good.



B. Manual Evaluation

We can recruit human raters to evaluate MT results. Here is an example: *Report by the Automatic Language Processing Advisory Committee* (1966).

When a translation sentence was being rated for intelligibility, it was rated without reference to the original. "Fidelity" was measured indirectly: the rater was asked to gather whatever meaning he could from the translation sentence and then evaluate the original sentence for its "informativeness" in relation to what he had understood from the translation sentence. Thus, a rating of the original sentence as "highly informative" relative to the translation sentence would imply that the latter was lacking in fidelity.

Scale of Intelligibility

9—Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.

8—Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily "corrected."

7—Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.

6—The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements.

Postediting could leave this in nearly acceptable form.

5—The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.

4—Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical word untranslated.

3—Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.

2—Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.

1—Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

Scale of Informativeness

9—Extremely informative. Makes "all the difference in the world" in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation.)

8—Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely.

7—(Between 6 and 8.)

6—Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader "on the right track" as to the meaning intended.

5—(Between 4 and 6.)

4—In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.

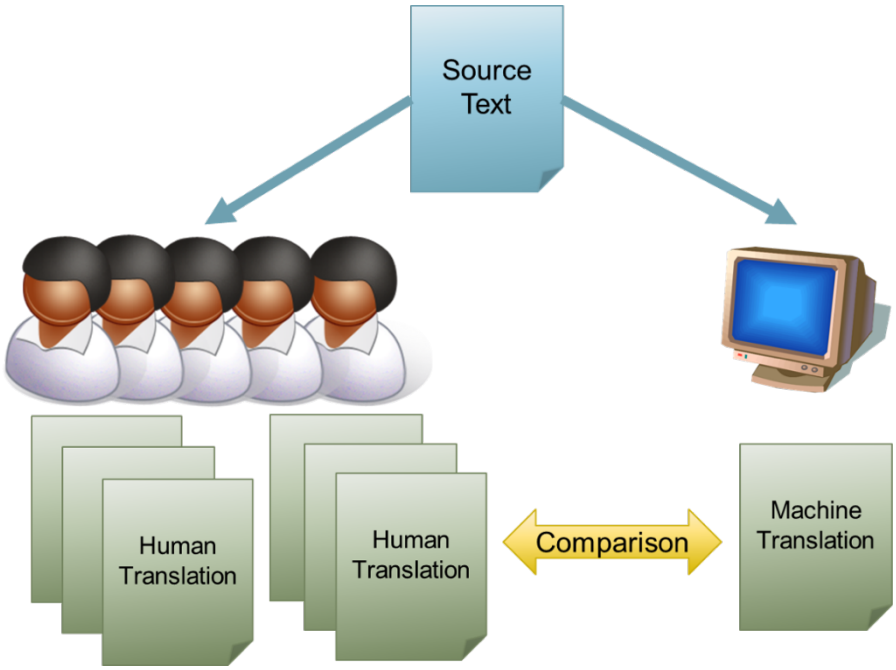
3—By correcting one or two possibly critical meanings, chiefly on the word

level, it gives a slightly different "twist" to the meaning conveyed by the translation. It adds no new information about sentence structure, however.
2—No really new meaning is added by the original, either at the word level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended.
1—Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced.
0—The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.

C. Automatic Evaluation

We can build automatic evaluation systems, which compare the similarity between the MT output and "model translations" by human translators. A high level of similarity suggests that the MT system is good.

Concept



Example 1: Precision

Given T = MT output and T_R = reference translation,

$$Precision = \frac{C(T \cap T_R)}{|T|}$$

Example

T = "Chinese officials responsibility of airport safety"

T_R = “Chinese officials are responsible for airport security”

$$Precision = \frac{3}{6} = 0.5$$

Example 2: Recall

Given T = MT output and T_R = reference translation,

$$Recall = \frac{C(T \cap T_R)}{|T_R|}$$

Example

T = “Chinese officials responsibility of airport safety”

T_R = “Chinese officials are responsible for airport security”

$$Recall = \frac{3}{7} = 0.43$$

Example 3: F-measure

$$F\text{-}measure = \frac{Precision \times Recall}{(Precision + Recall)/2}$$

Example

T = “Chinese officials responsibility of airport safety”

T_R = “Chinese officials are responsible for airport security”

$$F\text{-}measure = \frac{\frac{3}{6} \times \frac{3}{7}}{\left(\frac{3}{6} + \frac{3}{7}\right) \times \frac{1}{2}} = 0.46$$

Examples 4 and 5: BLEU and NIST

1. The BLEU (Bilingual Evaluation Understudy) score evaluates the quality of MT output by considering the number of n-grams that occur in both T and T_R. (Read more: http://aclweb.org/anthology/P/P02/P02-1040.pdf)
2. The NIST (National Institute of Standards and Technology) score is based on the BLEU score. The N-grams that are more informative (or occur less frequently) are given more weight. The bigram “business translation” is more informative than the bigram “of the” and should be given more weight. (Read more: http://www.mt-archive.info/HLT-2002-Doddington.pdf)