Odai Wardeh

TYPES OF EVALUATION CRITERIA FOR MACHINE TRANSLATION

The methodology of machine translation (MT) is still far from the semantic level or from processing ideas. There are misconceptions about the methodology of "deep neural networks" enhanced by "deep learning" because many people conceive the word "deep" as profound, while, the meaning of *deep* in this context comes simply from the fact that these neural networks have more layers than older networks. Anyhow, we have two types of evaluating MT systems: 1) human evaluation; 2) automatic evaluation.

Human evaluation of machine translation quality goes back many years. There are different types of human evaluation of MT, including 1) t*ypological evaluation*, which addresses translational phenomena the can be handled by a particular machine translation system; 2) *declarative evaluation*, which addresses how an MT system performs relative to various dimensions of translation quality; 3) *operational evaluation*, which establishes how effective a machine translation system is likely to be (in terms of cost) as part of a given translation process.

The main purpose of automatic evaluation is to establish objective metrics to assess MT outputs, which could be more reliable than the subjective estimations of translators. There are three assumptions that support automatic evaluation methods:

- the Reference Proximity Assumption (RPA);
- the Accuracy Assumption (ACA);
- the Human Likeness Assumption (HLA).

The Reference Proximity Assumption (RPA)

The human translation of the original is the quality reference to evaluate the machine translation. The quality degree is expressed with a metric obtained by an objective method, the distance between the machine translation, called hypothesis, and the human translation, called reference.

The Accuracy Assumption (ACA)

Evaluating sentence accuracy is not new as we have seen in human evaluations. The novelty is the automatic calculation of semantic similarities between machine translations and references.

The Human Likeness Assumption (HLA)

According to the HLA, a machine translation that resembles a human translation is good with the following: a) human/non-human translation classifier: the strategy turns evaluation into an automatic classification problem; b) Human Likeness and combination of the Reference Proximity Assumption measures as a meta-evaluation criterion that captures syntactic improvements, which are not captured by any single RPA measure. Their proposal is to combine the Reference Proximity Assumption measures that are good to distinguish machine translations and human translations in one metric.