Panel Perspectives on Machine Translation Evaluation

Moderator: Keith J. Miller, The MITRE Corporation

As one of the first applications of Human Language Technology, Machine Translation remains one of the most elusive. Although we can agree that we have made great strides in MT since the first systems were developed, these advances remain difficult to quantify. Indeed, over the years there have been many proposals for measuring the quality of Machine Translation, and various approaches to MT Evaluation have gained and lost popularity over time. Questions posed to panel participants will include:

- What can we see by looking at the development of machine translation evaluation alongside the development of machine translation itself over time?

- Are some types of evaluation better suited to certain audiences/uses? That is, are certain audiences better served by particular types of evaluation?

- Have we learned anything over the years of evaluation of machine translation, and if so, what?

- How can we leverage and best apply all that we have learned about MT evaluation in order to:
  - inform potential users/purchasers of MT?
  - help users to select the best MT technology for their particular use case(s)?
  - help to drive the state-of-the-art/quality of MT?

- Has MT evaluation reached the level of maturity of evaluation for human language technologies that appeared after it (e.g. Information Retrieval, Information Extraction, Language Identification, and others...)? If not, why not?

This panel will draw on experts in both human and automated metrics for MT evaluation from the EU, US and Asia. Short statements by the panelists, indicative of the types of discussions we are likely to have during the panel are below, followed by a short biographical statement of each panelist's involvement with Machine Translation and Machine Translation Evaluation:

Statements from Panelists:

Susan Armstrong, ISSCO, University of Geneva

MT evaluation and evaluation of translation technology in general has many facets, where performance quality vis-à-vis a given task, is but one. In the deployment of such technology, e.g. in a translation service, a larger context must be considered. The use of the technology to produce translations of a certain quality must take into account the process, the users and the uses, in other words, the scenario(s) in which the technology will be deployed. Since development of technology such as MT or translation memory software is not an end in itself, the context in which it will be used plays an important role in the evaluation procedure. In this context we consider the different factors that can influence the evaluation model from choice and adequacy of technology, availability and selection of data sets for development and deployment,

typology of different users (e.g. translators as post editors or the end users) and different uses (e.g. gisting vs. published translations). The current demand for improved technology to meet global translation needs calls for an investigation of evaluation in this larger context.

Gregor Leusch, RWTH Aachen University
When looking at work on MT evaluation over the last years, a couple of new evaluation measures have been proposed. Some of them were mostly anecdotally motivated, for others correlation with human judgment has been used as evidence.
I am worried that we might be confusing correlation and causality here. In other words, everybody seems to consider the following three statements to be equivalent (where "c" is an automatic evaluation measure, and A and B are MT systems):

1) "c" correlates well with "human evaluation" on existing data
2) If A is better than B, then c(A) is better than c(B)
3) If c(A) is better than c(B), then A is better than B

Or even more extreme, people seem to assume that from

1a) I take a couple of arbitrary static MT systems, and c correlates well with some human evaluation measure on those (after tuning c on this data, of course) follows

3a) If I tune my system to get a good c, I get good translations.

I do not think that this is a valid conclusion, and therefore propose to look at simple, untuned evaluation measures, where we are at least able to understand the caveats, gaming opportunities, etc.

John White, The MITRE Corporation
Progress in MT evaluation over the last decade, more or less, has been encouraging. On balance. But that particular balance, like my bathroom scale, seems to head readily in the direction I do not want it to go at the slightest provocation, while yielding, say, two-tenths of a point after much tedious work and deprivation. In those happy circumstances I rush to create a before and after picture portraying the facts pertinent to the comparison of my old loathsome, hedonous self to the new, efficient, informative, intelligible Adonis that I have become as a result of improving by those two-tenths of a point.

The problem is that the before and after pictures look just alike to anybody who doesn't look at me every day. And so it is with the automatic MT scoring methodologies in wide acceptance today. It is a really good thing to have a scale that you can use by just stepping on it; this ensures that you will be expected to do it with some frequency. It's a good habit for researchers to get into. But the correspondence between some MT at some BLEU score and the eventual usefulness of that MT for something sometime somewhere -- well, that is like when I interpret the bad news on the bathroom scale into water weight or big bones or some quantum anomaly in the springs of the scale itself. The easy automatic measurement should directly predict some fit with a performance requirement; the number on the scale should tell me whether I can fit in my clothes today. And it doesn't. But we should keep on stepping on the scale; it could get better.

Billy Wong, City University of Hong Kong
The new methods of automatic MT evaluation in recent years have enhanced the efficiency of the evaluation practice. Such advancement on methodology, however, appears to be far head than theoretical development. For example, it remains unclear that on which aspect an MT system is deemed better than another one given a higher score of automatic metric. Apart from the statistical evidence showing the correlation between the judgments of automatic and human evaluation, further theoretical account bridging the two sides is necessary.

Biographical Statements of Panelists:

Susan Armstrong has been working in the field of NLP for over 25 years when she first joined ISSCO, one of the oldest, internationally renowned institutes in the field of MT and evaluation, among others. She participated actively in the first data collection initiatives and projects in the exploitation of corpora, bringing the interest of this now well established field to the CL community. Currently, as a professor of translation technology at the University of Geneva, she has continued to pursue her interest in multilingual corpora and the use of the data to develop and improve translation tools. As consultant to translation services in international organizations, evaluation has been an important topic in order to best satisfy user needs in the context of internal resources and processes, available data and existing NLP technology.

Gregor Leusch received the degree in computer science from RWTH Aachen University, Aachen, Germany, in 2005, having written his diploma thesis on automatic evaluation measures for machine translation. He is currently pursuing the Ph.D. degree in the Department for Computer Science, RWTH Aachen University. His current research topics are the automatic evaluation and system combination for machine translation. The two topics share, in his view, a remarkably similar set of challenges compared to their speech recognition counterparts.
He is author or coauthor of several reviewed publications in international conferences and journals.

John White is a specialist in the field of machine translation evaluation with the MITRE Corporation. He supports the National Virtual Translation Center in meeting its goals of deploying the optimum configurations of translation technology for enhancing the workflow of a distributed human translation environment.
John has directed programs in the human language technology research and development disciplines since the 1970's, and has had the opportunity to work with government, industry, users, academics and researchers in most of the relevant sub-disciplines over the years.

Billy TM Wong is a PhD candidate in the Department of Chinese, Translation and Linguistics at the City University of Hong Kong. His major research focus is on MT evaluation, in particular its automated metrics and theoretical studies of the relationship between human and automatic measures. On this basis, the study on proper selection and use of MT technology for lay users is also explored.