

Machine Translation Quality Estimation: Lessons from a Typical NLP Task

Anil Kumar Singh
IIT (BHU), Varanasi, India
nlprnd@gmail.com

Abstract

As machine translation (MT) systems mature and are used for practical purposes, automatic quality estimation (QE) of their output acquires increasing importance. We present an outline of the work on MTQE and derive some conclusions. MTQE starts where the MT system has finished. It is the estimation of the quality of output of an MT system and it has to go beyond what the MT system has already considered, which is usually quite a lot. Researchers have been trying all kinds of information and techniques to get good quality estimates. Quality is something that even human beings do not always agree about. The degree of disagreement is remarkable, and it naturally makes the problem harder.

Like any NLP problem, MTQE has many aspects and is related to other problems. At present, it is at a stage where deployability is still a challenge. We discuss, how it being a typical NLP problem, the insights gained from it can often be easily transferred to other problems, and vice-versa. MTQE requires both statistical techniques and linguistic knowledge in some form. Learning to do MTQE well may be a good starting point for those who are just stepping in to the area of NLP.

Topical Outline

- Introduction: What is MTQE and why is it needed?
- A Typical NLP Problem: The usual scenario from NLP
- Multiple Formulations: Is it one problem or more?
- The State of the Art: A quick literary survey
- There is a Difference: What is special about MTQE?
- The Sources of Information: Selecting the features
- The Use of the Information: Implementation of feature engineering
- The Core Source: Be careful about the dataset
- Actual Estimation: Machine learning and inference
- As Expected: The architecture for solving many NLP problems
- A Part of the MT System: The problems of deployability
- The Experience: Some lessons learnt
- Summing Up: Some concluding remarks

Description of the Proposer

The proposer is a researcher (and currently a teacher) who has been working in the area of NLP for the last eleven years. He is working as a Visiting Faculty in the department of Computer Science and Engineering at IIT (BHU), Varanasi, India. He has published on various topics in NLP and has organized a couple of research workshops and a couple of introductory workshops on NLP. He did his PhD in Computational Linguistics from IIIT, Hyderabad, India. He is interested as much in development or implementation as he is on research (as his email id indicates). He is the creator of Sanchay, a collection of tools and APIs familiar to some researchers in India. He spent one year as a post-doctoral researcher at LIMSI-CNRS, Orsay, France. It was there that he worked on the topic of machine translation quality estimation (MTQE). The work there resulted in some publications and, equally importantly for him, in the development of a tool called Questimate (along with a companion tool called LatticeFst). This tool implements the complete workflow for MTQE, from feature extraction and selection to machine learning (with the help of Weka as a library) to analysis and visualization of the results. He has been associated over the years with machine translation and related activities in several different capacities. He also happens to be a professional translator and has translated from a wide variety of domains (from computers to poetry). The proposed tutorial is based to a large extent on his own experiences with MTQE.