Abstract

Since beginning, research in Machine Translation Evaluation has been a prominent area of study, with initial MT systems being evaluated by only human evaluators. This thesis is focused on studying various methods incorporated in MT Evaluation. A lot of evaluation campaigns have been undertaken for English as a target language, but evaluation of English-Hindi MT has not been taken up so well. Thus, we have focused our evaluations for this language pair. The basic motivation for undertaking this study was to help system managers of MT projects, as many times they have to wait for human evaluations to complete their assessments for an MT system. This process takes days to finish, which hinders the development process. Using the methods suggested in this thesis, a system manager can very easily analyze the performance of his system which can help him in keeping up with deadlines. We have developed a human evaluation metric which provides at par results with popular human adequacy and fluency metrics and further provides better answers as to why certain translations are better or worse. We have also studied various automatic evaluation metrics available across linguistic levels and tried to correlate the results of these metrics with human evaluation. For this we have used both, single as well as multiple reference translations. Further, we have used a combination of these metrics to get better correlations with human evaluations. We have also performed statistical significance testing on the results produced by automatic evaluation metrics to verify if the results produced by the metrics are really valid or are they good by just a matter of chance. At times it gets very difficult to get reference translations. Without reference translations no metric can provide evaluation results. Thus, we have also looked at measures to evaluate MT system without human references. For this we incorporated a trigram language model to evaluate the translations by providing ranks to outputs of various MT engines. We also developed two classifiers for this study. We further compared the results of these studies with human evaluations. We found that the results of these evaluations are at par with human evaluation.