

Predicting Sentence Translation Quality Using Extrinsic and Language Independent Features

Ergun Biçici · Declan Groves · Josef van Genabith

Received: date / Accepted: date

Abstract We develop a top performing model for automatic, accurate, and language independent prediction of sentence-level statistical machine translation (SMT) quality with or without looking at the translation outputs. We derive various feature functions measuring the closeness of a given test sentence to the training data and the difficulty of translating the sentence. We describe *mono* feature functions that are based on statistics of only one side of the parallel training corpora and *duo* feature functions that incorporate statistics involving both source and target sides of the training data. Overall, we describe novel, language independent, and SMT system extrinsic features for predicting the SMT performance, which also rank high during feature ranking evaluations.

We experiment with different learning settings, with or without looking at the translations, which help differentiate the contribution of different feature sets. We apply partial least squares and feature subset selection, both of which improve the results and we present ranking of the top features selected for each learning setting, providing an exhaustive analysis of the extrinsic features used. We show that by just looking at the test source sentences and not using the translation outputs at all, we can achieve better performance than a baseline system using SMT model dependent features that generated the translations. Furthermore, our prediction system is able to achieve the 2nd best performance overall according to the official results of the Quality Estimation Task (QET) challenge when also looking at the translation outputs. Our representation and features achieve the top performance in QET among the models using the SVR learning model.

1 Predicting Sentence Translation Quality

We investigate the prediction of sentence translation quality for machine translation (MT) without training a statistical MT (SMT) model or using the reference

Ergun Biçici · Declan Groves · Josef van Genabith
Centre for Next Generation Localisation,
Dublin City University, Dublin, Ireland.
E-mail: {ebicici, dgroves, josef}@computing.dcu.ie