

Bleu to the extent that it currently is. In this paper we give a number of counterexamples for Bleu's correlation with human judgments. We show that under some circumstances an improvement in Bleu is *not sufficient* to reflect a genuine improvement in translation quality, and in other circumstances that it is *not necessary* to improve Bleu in order to achieve a noticeable improvement in translation quality.