Towards better dialogue system evaluation

Yvette Graham
School of Computer Science and Statistics
Trinity College Dublin
ygraham@tcd.ie

Evaluation of open-domain dialogue systems is highly challenging, and development of better techniques is highlighted time and again as desperately needed. Despite substantial efforts to carry out reliable live evaluation of systems in recent competitions, annotations have been abandoned and reported as too unreliable to yield sensible results. This is a serious problem since automatic metrics are not known to provide a good indication of what may or may not be a high-quality conversation. Answering the distress call of competitions that have emphasized the urgent need for better evaluation techniques in dialogue, this talk presents the successful development of human evaluation that is highly reliable while still remaining feasible and low cost. Self-replication experiments reveal almost perfectly repeatable results with a correlation of r = 0.969.

Due to the lack of appropriate methods of statistical significance testing, the likelihood of potential improvements to systems occurring due to chance is rarely taken into account in dialogue evaluation, and the evaluation presented facilitates application of standard tests. Highly reliable evaluation methods then provide new insight into system performance and this talk includes a comparison of state-of-the-art models (i) with and without personas, to measure the contribution of personas to conversation quality, as well as (ii) prescribed versus freely chosen topics. Interestingly with respect to personas, results indicate that personas do not positively contribute to conversation quality as expected, a surprising result that will hopefully inspire discussion within the dialogue community.