複数の評価基準による対話システムの自動評価手法に関する研究

ロドリゲス　海

近年対話システムの実用化が増えている．対話システムは大きく分けて，対話を通してなんらかのタスクの達成を目的とするものと雑談自体を目的としたものに分けることができる．このうちタスクの達成を目的とするものはその達成率などからシステムの評価を行えるが，その一方で雑談対話における評価は明確な正解となる応答が定められておらず，評価が難しくなっている．また現状として，人手による評価が行われているが，これらの評価にはかなりのコストがかかってしまう．そこでこれらのコスト削減と雑談対話システムのパフォーマンスを測るための評価を自動化する方法が研究されている．

雑談対話の自動評価においてよく用いられる手法として正解とのシステムの応答の間の類似度を測る手法があるが，この手法では正解とできる応答が多岐に渡ってしまうことから自動評価による評価と人手による評価との間の相関をみたときに低くなってしまう問題があった．類似度を用いた自動評価手法の他に, 評価モデルの学習に基づく自動評価手法である Automatic Dialogue Evaluation Model (ADEM) がある。この手法では人手による評価との間で強く相関する評価ができるようになった．

これらの自動評価手法では現状として「対話の自然さ」という基準で評価が行われている．応答が対話の中で自然であれば，それは良い応答として評価される．しかし対話の評価基準となりうるものは「自然さ」以外にも存在するため，良い応答の中においても優劣をつけることができる．このような優劣を踏まえた評価が行えば，より適切な評価になると考えられる．本研究では対話システムの応答から会話をさらに発展させることができるものをより質の高い応答とした上でこの評価基準を「応答の継続性」とし,「応答の自然さ」という基準に加えて評価を行うことで，複数の基準軸で自動評価を行う必要性について検討する．また ADEM による自動評価手法において別の評価基準においてもこれまでの評価軸と同じように自動評価を行えるのかを検証する．

具体的には人手で生成した応答と対話モデルを用いて生成した応答を「応答の自然さ」と「応答の継続性」という二つの基準から評価してもらい，それぞれの基準による評価スコアの相関を測ることで基準の違いがスコアにどれだけ関連しているのかを調べた．またこれらの評価データを用いて ADEM を学習させ， ADEM が出力した評価スコアを比較することで, 評価基準が異なっても ADEM が同等のパフォーマンスを発揮できるのかを調べた．ここでのパフォーマンスとは ADEM の出力スコアと実際に人がつけたスコアの相関を取ることで，人に近いスコアをつけるものができたものをパフォーマンスが高いものとした．

本研究の結果として，対話とその応答に対して二つの基準で評価を行ったところ，応答文の質が高いほど評価基準によってその評価値に違いが生まれやすいことが分かった．また ADEM において「対話の継続性」での評価データから学習した際に,「対話の自然さ」での評価データから学習したものと比べて少しばかりパフォーマンスが劣る結果となった．

（指導教員　若林 啓）