人工知能学会研究会資料
SIG-SLUD-B401-01

大規模マルチリファレンスに基づく雑談対話システムの自動評価に向けた実験的検討

Experimental Analysis for Automatic Evaluation of Open-domain Conversational Systems based on Large-scale Multi-references

杉山 弘晃 [1*]　目黒 豊美 [1]　東中 竜一郎 [1,2]
Hiroaki Sugiyama [1]　Toyomi Meguro [1]　Ryuichiro Higashinaka [1,2]

[1] NTT コミュニケーション科学基礎研究所
[1] NTT Communication Science Laboratories
[2] NTT メディアインテリジェンス研究所
[2] NTT Media Intelligence Laboratories

Abstract:
The evaluation of conversational systems that chat with people remains an open-problem. Some studies have evaluated them by hand with ordinal scales like the Likert scale. One limitation with this approach is that we cannot use the previously evaluated values since the ordinal scales are not consistent across all of the evaluations. This makes it difficult to compare proposed and previous systems since we have to implement the previous systems and simultaneously evaluate them. We propose an automatic evaluation method for conversational systems that evaluates the sentences generated by systems on the basis of the similarities that are calculated with many reference sentences and their annotated evaluation values. Our proposed method's correlation coefficient with humans reached 0.514, and that of the human annotators was 0.783. Although there remains a gap between the estimated and the human-annotated values, the proposed method outperforms a baseline method that uses the BLEU scores as the evaluation values. We also show that we can gain a correlation coefficient of 0.499 with evaluating just 7% of all the data.

1　序論

近年，従来のタスク指向の対話システムとは異なる，エンタテインメントやカウンセリングなどを目的とした，雑談を行う対話システムに注目が集まっている [大西 14, Ritter 11, Wong 12]．雑談対話システム研究を進める上での課題の一つが，構築したシステムの評価である．システムを簡便に評価する仕組みは，システムを逐次的に改善していく上で不可欠な要素である．

タスクの遂行を目的とする対話システムでは，タスクの達成率や達成にかかる時間などの明確な評価基準があるため，システムの評価は比較的容易である [Janarthanam 08, Walker 97]．一方，雑談対話システムでは，システムが出力すべき正解が必ずしも自明ではない．そのため従来は，システムの出力文に対し人手で Likert 尺度などの順序尺度の評価値を付与し，平均値をとる方法が主流であった（e.g., [稲葉 14]）．しかし，順序尺度で付与される値は相対値であるため，順序関係は一貫性があるものの，評価毎に平均値は異なる可能性がある．そのため，従来システムと提案システムを付与された評価値の平均値で比較するには，比較対象の従来システムを再実装し，同一の評価者が値の一貫性を保って，提案システムと同時に評価する必要がある．この方法は非常にコストがかかるため，既存研究との比較は容易ではない．この課題を解決するには，再現可能な形で，自動的に評価値を付与できる仕組みが必要である．

システムが出力する文を評価する枠組みとして，機械翻訳の分野では，ある入力文に対してシステムが出力した文をリファレンス文と比較することで評価値を推定する，BLEU や ROUGE などの手法が知られている [Papineni 02, Isozaki 10]．Ritter らや長谷川らは，リファレンス文を 1 文のみ用いた BLEU を評価値としてシステムごとの平均値を計算し，人手評価と比較している [Ritter 11, 長谷 14]．しかしながら，機械翻訳に比べて正解とすべき出力文の範囲が広い雑談対話では，1 文のリファレンス文のみでは正解とすべき文の範囲をカバーできず，推定評価値が人手評価と大きく異な

*連絡先：NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
E-mail: sugiyama.hiroaki@lab.ntt.co.jp

る場合があった．

本研究ではこの問題を解決するため，リファレンス文として正例を大規模に収集するとともに，負例となりうる文を追加し，かつこれらに評価値を付与するアプローチを提案する．負例の追加と評価値の付与により，単純に正例を増やすよりも正例の範囲をより明確にでき，システム評価の推定精度を向上できると考えられる．本研究では，リファレンス文の収集方法を説明するとともに，リファレンス文の大規模化および評価値付与によって，推定評価値と人手で付与した評価値の相関がどの程度向上するかについて報告する．

2 関連研究

タスク対話システムを自動的に評価する試みとして，Walker らが提案した PARADISE がある [Walker 97]．これは，システムと人の間でなされた対話に対し，対話から得られる発話文の長さや発言数などの特徴量に基づいて，その対話の質を評価する方法である．この方法は観測された対話を評価する手法であるため，対話システムを評価するためには，その都度対話を行う必要がある．ここで対話相手を人とした場合，人ごとに嗜好や口調が異なる上，対話をする度に対話に対する興味が変動するため，得られる対話の質が実験の都度ばらつくという問題が生じる．相手を別の対話システムとすると，応答パターンの固定化はしやすいものの，現時点で人と同様の質を保って応答できるシステムが存在しないことから，対話の質が対話相手となるシステムによって悪化するおそれがある． そのため，実際の対話を介さない形式での評価が望ましい．

Devault らや Gandhe らは，ロールプレイを行う対話システムに対して，発話履歴中の単語を特徴量とした発話文の自動評価を行っており，人手評価に対して 0.48 程度の相関を得ている [Devault 11, Gandhe 14]．しかしながら，彼らの実験では，システムが出力できる発話は 96 種類のあらかじめ作成した文に限定されており，任意の文が出力される対話システムでの有用性は検証されていない．

システムが出力する任意の文を自動的に評価する手法は，機械翻訳の分野において盛んに研究されている．特に，入力文に対してシステムが出力した文とリファレンス文との類似度を単語 N-gram の一致率などに基づいて計算し，その値を評価値とする BLEU や ROUGE などの自動評価尺度が知られている [Papineni 02, Isozaki 10]．これらの値は，機械翻訳において人手による評価値と非常に高い相関値を示すことが知られており，機械翻訳システムの性能を評価する有力な尺度となっている．Ritter らや長谷川らは，BLEU を用いて，システムの発話の適切さの自動評価を試みている [Ritter 11, 長谷 14]．彼らは，複数の雑談対話システムが出力した応答文のペアについて優劣を評価し，優劣を基に各システムの勝率を得るとともに，システムごとに応答文の BLEU の平均値を計算して，各システムペアについて勝率の大小と BLEU の大小が一致するかを調べた．Ritter らは，6 ペア中 5 ペアが一致していたと報告しており，ある程度人手の評価を反映してはいるが，そのまま人手評価の代替と用いるのは難しいとしている．彼らの報告において BLEU があまり有効でなかった理由として，彼らがリファレンス文を 1 文のみ用いていたことが挙げられる．BLEU のような距離に基づく手法で評価値を推定するには，ある程度正例となる文をカバーしていることが求められる．しかし，雑談対話と機械翻訳を比較すると，雑談対話では機械翻訳にくらべ正例とされる出力文の範囲が広いため，数文のリファレンス文では正解とすべき文の範囲をカバーしきれず，適切に文を評価できなかったものと考えられる．

3 リファレンス文の作成

本研究では，雑談対話システムが入力文に対して出力する応答文の適切さを，リファレンス文との距離を利用して推定する．そのためには，リファレンス文が，正解とされる応答文（正例）の範囲をカバーできている必要がある．本研究ではまず，リファレンス文の数を大規模化することで正例のカバー率の向上を目指す．しかし，単純にリファレンス文の数を増やしても，雑談対話システムでは正例の範囲が非常に広いため，正例の範囲をカバーしきれるとは限らない．また，どの程度まで正例かがわからないという問題もある．加えて，システムの性能が人に比べて低い場合，システムが出力する文はいずれのリファレンス文とも距離が遠くなる可能性があり，こうした文ばかりをシステムが出力する場合，システム間で評価値に差がつかず，適切に優劣を評価出来なくなる恐れがある．

そこで本研究では，負例となる文をリファレンス文に加えるとともに，得られたリファレンス文に評価値を付与するアプローチを提案する．負例の追加と評価値の付与により，単純に正例を増やすよりも正例の範囲をより明確にでき，システム評価の推定精度を向上できると考えられる．

3.1 入力文の収集とリファレンス文の作成

まず，システムへ入力するための入力文を収集する．入力文は，文単体で何についての発話であるかが理解できるように書かれている必要がある．そのため本研究では，Web や対話実験ログから人が記述した発話文を収集し，これらに対して理解しやすさ (了解性) を人手で付与することで，文単体で理解しやすい，了解性が高い文を集める．

次に，各入力文に対し，複数のリファレンス文作成者が文を作成する．このとき，負例となるリファレンス文を得るため，一部の文作成に制約を加える本研究では，文作成時の制約として，リファレンス文の文字

数制約と，入力文のマスクを行う．文字数を制約することで，使える表現が制約され，リファレンス文に不自然な表現が含まれる効果が期待できる．また，入力文の一部を隠して作成者へ提示することで，入力文と話題が異なる文が得られると考えられる．

人手で作成したリファレンス文に加えて，検索ベース対話システムやルールベース対話システムなど，既存の対話システムから得られた発話をリファレンス文に加える．現在の対話システムは必ずしも適切な応答を返せていないため，負例と正例が適度に混在した文が得られると予想される．

3.2 評価値の付与

作成したリファレンス文集合に対し，人手で評価値を付与する．本研究では,「応答文としての自然さ」を評価基準として文のペアごとに優劣を人手で評価し，全てのリファレンス文に対する勝率をある文の評価値とする．すなわち，全てのリファレンス文に対して自然であると判断されたリファレンス文は評価値として 1 が付与され，逆に全てのリファレンス文よりも不自然であると判断された文は 0 が付与される．Likert 尺度のような順序尺度を付与しない理由は，予備実験として 7 段階の Likert 尺度で評価値を付与したところ，約 45%が最大値の 7，約 25%が最低値の 1 と分類され，リファレンス文間の違いを得にくかったためである．ペアワイズで比較する場合，$N(N-1)/2$ 回比較する必要があり，直接順序尺度で評価値を付与するよりもコストが大きくなるというデメリットがある．しかし，ペアワイズの優劣から得られた勝率は 0 から 1 の間で満遍なく分布しているため，最低値の 1 と評価されるものの中でも優劣を付けられるという利点がある．また，順序尺度評価では，アノテータ間で評価値に対する感覚が異なるため，アノテータ間で値を直接比較できないという問題があるが，勝率のような比率尺度を用いることで，こうした感覚の差異を吸収できる利点もある．さらに，Sculley は，こうしたペアワイズは必ずしも全ペアに対し行う必要はなく，一部をサンプリングしてもあまり精度に影響はないと報告している [Sculley 09]．これが本研究の対話システムの自動評価にもあてはまれば，アノテーションコストを抑えつつ，上記の利点を活かすことが可能になると考えられる．

4 評価値の推定

本研究では，雑談対話システムが出力する応答文の評価値を，大規模な評価値付きのリファレンス文集合を用いて推定する．評価値付きのリファレンスを用いる方法として，以下の 4 種類の方法が考えられる．

類似度平均　システム出力文と各リファレンス文の間で，機械翻訳の自動評価で用いられる BLEU[Papineni 02] や RIBES[Isozaki 10]，WER(Word Error Rate) などの，文間の類似度を表す尺度を計算し，上位 N 個の平均値をシステム出力文の評価値とする方法である．評価値を用いずリファレンス文との類似度のみを用いている点で，従来の機械翻訳に最も近い方法である．評価の低いリファレンス文が含まれている場合，それらと類似度が高い場合も推定評価値が高くなるため，本研究では，入力文にマスクをして作成したリファレンス文と，システムが出力したリファレンス文をリファレンス文集合から取り除いて平均値を計算する．

評価値平均　類似度の大きさが上位 N 個の評価値の平均値を，システム出力文の評価値とする方法である．類似度は 1 つ目の類似度平均と同じものを用いる．1 つ目の方法に比べ，直接的に評価値を利用するため，特に評価の低いリファレンスとの類似度が大きい場合に，適切に低い評価値を付与できると予想される．

評価値重み付け　類似度の大きさが上位 N 個の評価値を類似度で重み付けて平均した値を，システム出力文の評価値とする方法である．2 つ目の方法に比べ，類似度をより強く評価に反映できると考えられる．

回帰モデル　SVR[Smola 04] などの回帰モデルを用い，各リファレンス文に対する BLEU などの類似度を特徴量として評価値を推定する方法である．本研究では，入力文ごとに回帰モデルを学習する．

5 実験

本研究では，収集された文の評価値の分布や人同士の相関について分析すると共に，収集したリファレンス文の評価値を leave-one-out 法で推定し，正解値との相関を調べることで，提案手法の有効性を検証する．

5.1 実験設定

5.1.1 入力文の収集

入力文を収集するコーパスとして，本研究では，我々が収集した雑談対話コーパス [Higashinaka 14] と Twitter コーパスを用いる．雑談対話コーパスは，のべ 360 名以上の話者から，1 対 1 のテキストチャット形式による雑談を，計 3680 対話，約 13 万文収集したものである．これに，目黒ら [Meguro 10] によって定義された対話行為を付与し，自己開示（自分についての事実や経験などを話した発話），質問，もしくは情報提供に関する対話行為が付与された文を，入力文の候補として抽出した．一方，Twitter コーパスから了解性を高い文を容易に収集する方法として，話題となりうる単語を含む文を Twitter から検索し，そのうち非文でないものをルールで抽出する，稲葉らの方法 [稲葉 14] がある．本研究ではこれを参考に，話題となりうる単語（Google trends 2012 in Japan[1]の各カテゴリで 10 位以上の単語句のうち,「Xperia acro HD」などのよう

[1] https://www.google.co.jp/trends/topcharts#date=2012

に，空白を含まないもの）を含むおよそ 1 億 5 千万ツイートを入力文の候補として抽出した．収集した入力文候補について，筆者ら以外の 1 名のアノテータが，5 段階の Likert 尺度で内容の了解性を付与した．そのうち，最良値の 5（内容は省略なく明確に記述されている）を得た文から，コーパスごとに 5 文をランダムに選び入力文とした．

5.1.2 リファレンス文の作成

各入力文に対し，10 名のリファレンス文作成者が，自由に 3 文，10 文字以上の文を 3 文，10 文字未満の 1 文を 1 つ，計 7 文作成した．このとき，自由入力を含めて文字数は 50 文字以内とした．また，対話中の発話であることを意識し，話を続けたくなるように作成するように指示した．

本研究では，負例のリファレンス文を作成するため，上記の文字数制限に加え，入力文の一部を文節単位でマスクしてリファレンス文作成者へ提示する．例えば，「何か得意なものはありますか?」という入力文の文節の 60%をマスクする場合，「なにか *** *** ありますか?」のように作成者へ提示される．ここでは，マスクしないものを 6 つ，全体の 30%をマスクしたものを 2 つ，60%をマスクしたものを 2 つ用意した．これらをランダムに 10 名のリファレンス文作成者に割り当て，マスク 1 つあたり 1 つの文節が入ることと，そこを想像しながらリファレンス文を作成することを作成者に指示した．以上より，1 つの入力文に対し，マスク無しの 42 文，30%マスクの 14 文，60%マスクの 14 文の計 70 文が得られる．

さらに，人手で作成したリファレンス文に加えて，Ritter らが提案した検索ベースの発話生成手法である，IR-status，IR-response からそれぞれ 10 文，我々が開発したルールベース対話システムから 10 文収集した [Ritter 11, 目黒 14]．IR-status とは，Twitter から入力文に類似したツイート (status) を検索し，in-reply-to 機能で対応付けられた返信ツイート (response) をシステムの発話文として出力する手法である．IR-response は，Twitter 中の返信ツイート (response) から直接入力文に類似する文を検索する手法である．ルールベース対話システムは，入力文との一致を調べるパターンとそれに紐づいた出力文のペアを人手で記述したシステムである．パターンの検索には TF-IDF で重み付けた単語の cos 類似度を用い，類似度が高い 10 文をリファレンス文へ追加した．最終的に得られたリファレンス文集合は，1 入力文あたり，人手 70 文，検索ベース 20 文，ルールベース 10 文の計 100 文である．

5.1.3 評価値の付与

本研究では，得られた 10 個の入力文とリファレンス文集合のペアについて，2 名の評価者が評価を付与した．10 入力文のみを対象とした理由は，N が 100 と

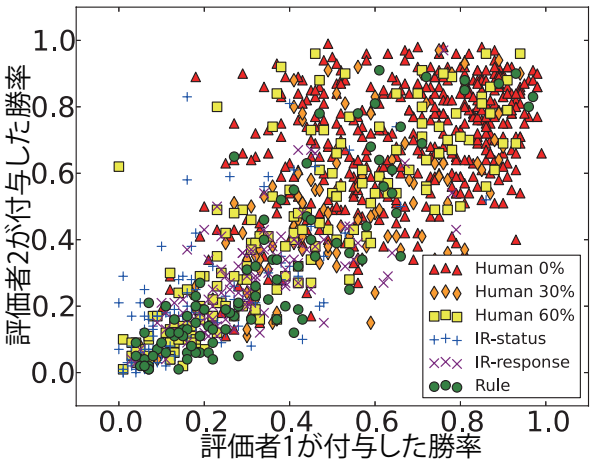

図 1: 評価者間の評価値の分布．赤い三角は人作成でマスク無し，橙色のダイヤは人作成でマスク 30%，黄色い四角は人作成でマスク 60%，青い十字は IR-status，紫色の×は IR-response，緑色の丸はルールベースから得られたリファレンス文を表す

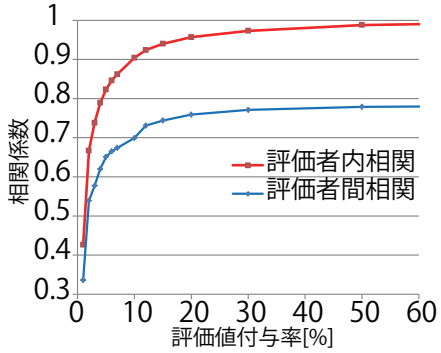

図 2: 評価付与数の変化に対する相関係数の変動

大きく，評価回数が 1 入力文につき 4950 回と膨大になってしまうためである．ただし，本研究では全てに対し評価を付与するが，一部をサンプリングして付与しても，精度にあまり影響しないことが知られている [Sculley 09]．本研究ではサンプリングに対する精度への影響を調べ，どの程度評価コストを下げられるかについても検証する．表 1 に，用いた入力文，およびリファレンス文の例とそれらの勝率を示す．

5.2 評価値の分析

まず，付与された評価値の分布と人同士の相関について調べる．図 1 に評価者間の勝率の分布を示す．勝率は 0 から 1 の範囲でおおよそ均等にばらついており，両端や中央に集中するような偏りはないことがわかる．全体的に人手で作成した文が高く評価され，システムが出力した文は低く評価されている．システムの中では，Twitter を用いた IR-status と IR-response は低〜中程度の評価に固まっている一方，ルールベースは高い評価にも分布しており，適切なルールに合致する場合は人に匹敵する応答を出力できていることがわかる．またこのとき，評価者間の勝率の相関は，ピアソンの積率相関係数を用いて計算すると，0.783 と比較的高い

表 1: 用いた入力文とリファレンス文，およびその勝率．リファレンス文は勝率最大と最小を利用

入力文	リファレンス文例	勝率
そして、ディズニーランドの大混雑も苦手です…。	開門と同時に皆が走り出す光景って、何度見てもぞっとしますよね。	0.96
	飼っている人がいると聞いてびっくりです。	0.01
なにかスポーツをされていますか？	毎日三十分くらいジョギングしてます	0.98
	清水義範さんとかどうですか。	0.01
LAWSON 寄ったらいきなり紅蓮の弓矢流れて噴いた	紅蓮の弓矢って、誰の曲ですか。	0.98
	www	0.0
ゴテゴテしいポッキーしか売ってないので別のコンビニ行こう	最近ごてごてポッキーが流行ってますね。	0.94
	あと五日我慢すれば君も地球防衛軍の仲間入りやで!(にっこり)	0.02
iTunes に入ってるの確認したらアニソンとゲーソンとボカロと声優さんとドラマ CD だらけだった	アニメとかゲームが好きなんだね！	0.95
	福袋なので仕方ないです。	0.03
何か得意なものはありますか？	パソコンの扱いが得意です	0.93
	(´・ω・`)	0.0
文字の攻撃力というものを理解しておかねばならない	人を傷つけることもあるということを理解しないといけませんね。	0.97
	おめ!	0.0
和菓子は食べられますか？	和菓子は大好きです	0.99
	風壮「お前、痩せたいんじゃないのかよ」	0.03
日本各地が安定した天気になるよう願いまして、おやすみ	私も願っています！	0.98
	青火効果ですね (*´ω`*)	0.01
自分も妹がいますが、気づかずに同じ漫画を買ってきちゃうことはあります。	保存用と読む用にするしかないですね。	0.98
	なんだ、眠いのか? それなら添い寝してやるから少しだけ寝ちまえよ。	0.01

相関を示していた．図 1 より，特に低い評価が付与された文の相関が強いことがわかる．

次に，図 2 に評価付与数を変化させた場合の，同一評価者内および評価者間の相関値の変化を示す．各付与率における各文の勝率は，付与されているペアワイズ評価をランダムにサンプリングして得る．図 2 より，おおよそ 12%程度を境に，評価者間・評価者内ともに増加が緩やかになっている．本実験の 100 リファレンス文の場合，入力文ごとに 600 ペアを比較することで，全ペア (4950 ペア) を比較する場合と強い相関 ($r = 0.924$) を持つ評価値を得ることができる．

5.3 評価値推定結果

本研究では，4 章に示す 4 手法を比較する．評価値の推定は leave-one-out 法で行った．各手法のパラメータは実験的に，最も良い性能を示したものを利用する．類似度平均，評価値重み付けの N を 3 とし，回帰モデルには RBF カーネルの SVR を $C = 5$ で用いた．また，各手法で用いる類似度として，BLEU，RIBES，WER を比較する．WER は Normalized Levenshtein 距離を用いて計算し，類似度として扱うための補正を加える．ここでは，$WER_{sim} = 1 - WER_{dis}$ で範囲を 0 から 1 とする場合と，$WER_{sim} = 1 - 2WER_{dis}$ で-1 から 1 とする場合の 2 通りを比較する．

図 3 に，各類似度と手法を組み合わせた場合の，相関係数を示す．最も高い相関係数を示したのは，WER を-1 から 1 の範囲として特徴量に用いた SVR で，相関係数は 0.514 であった．人同士の相関係数の 0.783 には有意に差があるものの，シンプルなアプローチで比較的強い相関が得られている．通常の機械翻訳と同

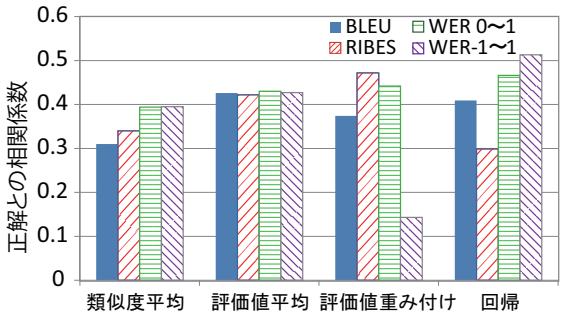

図 3: 推定された評価値と正解との相関

様に，評価値を用いず正例のみから推定する手法では，WER を類似度に用いた場合が BLEU，RIBES を用いた場合よりも高かった．しかしながら，相関係数は 0.4 程度に留まっており，他の評価値を用いた手法よりも低かった．これは，評価値の付与が推定精度の向上に有用であることを示している．また，評価値の重み付けでは RIBES を類似度に用いた場合が 0.472 と高い相関係数を示していた．評価値重み付けで BLEU よりも RIBES が有効に働いた理由は，BLEU では 1gram から 4gram までの単語一致率に基づいて計算する一方，RIBES では一致した単語の並び順に基づいて計算するため，BLEU が用いている高次の N-gram では一致が得られず，類似度に差がつきにくくなっていたためと考えられる．

図 4 に，上記設定で最も相関係数高かった，WER を用いた SVR について，ペアワイズ評価を付与する割合を変化させた時の相関係数の変化を示す．相関係数は評価割合が 7%に達するまでは急激に上昇し，7%時の相関係数は 0.499 を示した．実用上は 0.499 と中程度の相関でも有用な場合もあるため，そうした場合は全体の 7%（各入力文に対して 350 ペア）に評価を付与する

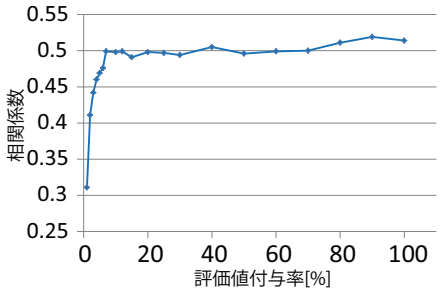

図 4: 評価付与割合に対する推定評価値と正解との相関

表 2: システム単位の平均評価値．H は人手評価を表し，後ろの数字はマスク率を表す

システム	H. 0%	H. 30%	H. 60%	IR-st.	IR-re.	Rule
人手評価	0.666	0.518	0.411	0.202	0.326	0.346
推定値	0.570	0.487	0.503	0.381	0.345	0.465

だけでシステムを構築できるといえる．その一方，上昇は 7%を境に止まり，その後 80%まで停滞していた．人同士では 12%まで線形に増加し，その後ゆるやかになるものの停滞はしなかったことと比較すると，今回用いた回帰モデルと特徴量では捉えきれない人の評価基準があるものと考えられる．

表 2 に，人が付与した評価値と SVR で推定した評価値を，システム単位で平均した結果を示す．おおまかには相関し，相関係数も 0.849 と高い値を示していた．しかし，15 ペア中，マスク率 30%での人手作成とマスク率 60%での人手作成の組み合わせ，および IR-status と IR-response の組み合わせの 2 ペアで，人手評価との逆転が見られた．人手評価と比較すると，マスク率 60%，IR-status の評価値が高く推定されている．マスク率 60%で評価値が低い場合は，応答する話題のみが誤っており，それ以外の言い回しは正しい場合が多いと予想される．単語の重要度を用いずに文間の類似度を計算する WER では，そうした違いを反映しきれないため，推定精度が低下したと考えられる．この問題は，TF-IDF などで単語に重みを与えることで改善できると考えられる．また，IR-status は，顔文字での応答や「www」など，他のシステムが出力するリファレンス文には現れない言い回しが多かった．これらとの文間の異なり具合は，WER や BLEU のような単語一致を前提とする類似度のみでは表現しきれなかったと考えられる．文字種の割合や，文字 N-gram 特徴を導入することで，表現可能になると考えられる．

6 結論

本研究では，雑談対話システムが出力する発話文の適切さを自動で評価する手法を提案し，その有効性を検証した．提案手法は，入力文に対するリファレンス文とその評価値を大規模に収集し，評価対象の文と各リファレンス文との WER 距離を特徴量とする回帰モデルを用いて，文の評価値を推定する手法である．人同士の評価値の相関係数が 0.783 であったのに対し，本手法は 0.514 と比較的強い相関を示した．また，リファレンス文の評価値を全体の 7%のみに付与した場合も，0.499 と全てに付与した場合とほぼ同程度の相関となることを示した．

展望として，複数ターンから成る対話実験で得られる評価との比較や，入力文数の拡充による検証の妥当性向上，言語特徴などの新しい特徴量の導入を進めたい．

参考文献

[Devault 11] Devault, D., Leuski, A., and Sagae, K.: Toward Learning and Evaluation of Dialogue Policies with Text Examples, in *Proc. SIGDIAL*, pp. 39–48 (2011)

[Gandhe 14] Gandhe, S. and Traum, D.: A semi-automated evaluation metric for dialogue model coherence, in *Proc. IWSDS*, pp. 141–150 (2014)

[Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proc. COLING*, pp. 928–939 (2014)

[Isozaki 10] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs, in *Proc. EMNLP*, pp. 944–952 (2010)

[Janarthanam 08] Janarthanam, S. and Lemon, O.: User simulations for online adaptation and knowledge-alignment in Troubleshooting dialogue systems, in *Proc. LONDIAL*, Vol. 45 (2008)

[Meguro 10] Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K.: Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes, in *Proc. COLING*, pp. 761–769 (2010)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.: BLEU: a method for automatic evaluation of machine translation, in *Proc. ACL*, No. July, pp. 311–318 (2002)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W.: Data-Driven Response Generation in Social Media, in *Proc. EMNLP*, pp. 583–593 (2011)

[Sculley 09] Sculley, D.: Large Scale Learning to Rank, in *NIPS 2009 Workshop on Advances in Ranking*, pp. 1–6 (2009)

[Smola 04] Smola, A. J. and Schölkopf, B.: A Tutorial on Support Vector Regression ‡, *Statistics and computing*, Vol. 14, No. 3, pp. 199–222 (2004)

[Walker 97] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents, in *Proc. EACL*, pp. 271–280 (1997)

[Wong 12] Wong, W., Cavedon, L., Thangarajah, J., and Padgham, L.: Strategies for Mixed-Initiative Conversation Management using Question-Answer Pairs, in *Proc. COLING*, pp. 2821–2834 (2012)

[稲葉 14] 稲葉通将, 神園彩香, 高橋健一：Twitter を用いた非タスク指向型対話システムのための発話候補文獲得, 人工知能学会論文誌, Vol. 29, No. 1, pp. 21–31 (2014)

[大西 14] 大西可奈子, 吉村健：コンピュータとの自然な会話を実現する雑談対話技術, NTT DoCoMo テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17–21 (2014)

[長谷 14] 長谷川貴之, 鍛治伸裕, 吉永直樹, 豊田正史：オンライン上の対話における聞き手の感情の予測と喚起, 人工知能学会論文誌, Vol. 29, No. 1, pp. 90–99 (2014)

[目黒 14] 目黒豊美, 杉山弘晃, 東中竜一郎, 南泰浩：ルールベース発話生成と統計的発話生成の融合に基づく対話システムの構築, 人工知能学会全国大会 (2014)