疑似応答を用いた雑談対話システムの自動評価

蔦 侑磨[1,a)]　吉永 直樹[2,b)]　豊田 正史[2,c)]

概要：雑談では発話に対して多様な内容・スタイルの応答が可能であるが，雑談対話システムの評価に人の会話データを利用する場合，参照応答としては基本的に特定の個人が行った一応答のみしか利用できないため，応答の多様性を考慮することが困難である．この問題に対し，入力発話-参照応答ペアに類似する発話-応答ペアの応答を疑似応答として大規模対話データなどから収集し，人手で応答としての妥当性を付与して評価に利用する評価手法 ΔBLEU が存在する．しかし，これをオープンドメインな雑談応答生成の評価に足るだけの大規模評価データの構築に用いることはコスト的に現実的でない．そこで本研究では，大規模対話データ中で複数応答を持つ発話から学習された分類器によって，疑似応答に対する妥当性付与と選別を行って ΔBLEU を自動化する ΔBLEU-auto を提案する．実験では大規模な Twitter データを利用して，Transformer に基づく雑談対話応答システムの評価を提案評価手法により評価した．その結果，提案評価手法により人手評価との相関に関して ΔBLEU と同等以上の相関が得られることを確認した．

1. はじめに

Apple Siri や Amazon Alexa, Google Assistant や LINE Clova など人と会話行う知的対話エージェントへの関心が高まりつつある．その流れを受けて，質問応答のような応答内容や目的が明確なタスク指向型対話だけでなく，雑談的な対話である非タスク指向型対話（以下，雑談対話）に関する研究 [6], [12], [13] が盛んに行われている．

雑談対話システムで中心的に研究される応答生成研究における主要課題として，生成応答に対する自動評価尺度が確立していないことが挙げられる．現状，雑談応答生成の評価において用いられている BLEU [10] や ROUGE [7] などの自動評価尺度は機械翻訳や要約などの雑談応答生成とは別のテキスト生成タスクから転用されたものであり，雑談応答生成の評価に用いた場合，人手評価との相関が低いことが問題として指摘されている [8]．これは，機械翻訳や自動要約に比べて雑談応答生成ではスタイル・内容共に多様な出力 (応答) が可能であるにも関わらず，雑談応答生成の評価に用いられる人の対話では，ほとんどの場合，参照応答として特定の個人が行った一応答のみしか利用できないためである．

この問題に対し，雑談対話の応答多様性を考慮した自動評価手法として ΔBLEU [3] が提案されている．この手法では評価対象の生成応答の応答元である入力発話に対して Twitter 上の大規模対話データセットから疑似応答を収集して参照応答に追加する Sordoni [14] らの手法を拡張し，各参照応答に応答としての妥当性を人手で付与して用いることで応答の多様性に対処している．ΔBLEU では人手評価と高い相関が得られているものの，この手法をオープンドメインな雑談応答生成の評価に足るだけの大規模評価データの構築に用いることはコスト的に現実的でない．

本研究ではこの問題を解決するために，Twitter から自動収集した疑似応答候補に対し，自動生成した教師データで学習した分類器によって応答妥当性の付与および選別を可能にする評価手法 ΔBLEU-auto を提案する．また，ΔBLEU では，入力発話-参照応答ペアと類似する発話-応答ペアの応答を疑似応答として収集しているが，応答の類似性まで考慮して疑似応答の収集を行うと，内容的に多様な応答の収集が難しくなる．そこで本研究では，疑似応答候補の収集の際に入力発話のみに類似する発話の応答を疑似応答候補として収集し，より多様な応答を収集することを試みる．

実験では著者の所属する研究室で継続的に収集を行っている大規模 Twitter アーカイブ上の英語対話データを利用して，提案評価手法で評価する雑談応答生成モデルの学習，疑似応答候補の収集，および疑似応答に自動で評価付与を行うための分類器の学習を行い，これらを用いて提案評価手法の評価を行った．実験結果において提案評価手法が，雑談応答生成の評価において ΔBLEU を超える人手評価との相関が得られることを確認した．

[1] 東京大学大学院情報理工学系研究科
[2] 東京大学生産技術研究所
[a)] tsuta@tkl.iis.u-tokyo.ac.jp
[b)] ynaga@iis.u-tokyo.ac.jp
[c)] toyoda@tkl.iis.u-tokyo.ac.jp

2. 事前知識

本章では提案手法の先行研究である ΔBLEU [3]，並びに ΔBLEU の基礎となる BLEU [10] について説明する．

2.1 BLEU

BLEU [10] は表層類似性に基づく機械翻訳システムの標準的自動評価尺度であり，システム出力と参照出力で重複する n-gram の出現回数に基づきシステム出力の評価値を算出する．評価値は具体的には，短すぎる出力に対するペナルティ BP (Brevity Penalty) と修正 n-gram 精度 p_n に関する幾何平均を用いて以下の式で計算される．

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_n \frac{1}{N} \log p_n\right) \tag{1}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } \eta > \rho \\ e^{(1-\rho/\eta)} & \text{otherwise} \end{cases} \tag{2}$$

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{\#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \#_g(h_i)} \tag{3}$$

η, ρ はそれぞれ出力と参照出力の平均文長, n は n-gram の語数，$\{r_{i,j}\}$ と h_i は i 番目の入力に対する J 個の参照出力とシステム出力，$\#_g(u)$ は文 u における n -gram g の出現頻度，$\#_g(u,v)$ は $\min\{\#_g(u), \#_g(u)\}$ を意味する．

BLEU を雑談応答生成の評価に用いる場合，雑談対話では内容・スタイル共に多様な出力（応答）が可能であるにも関わらず，雑談応答生成の評価に用いられる人の対話では，基本的に参照応答として特定の個人が行った一応答のみしか利用できないなどの要因から，人手評価との相関が出ないことが指摘されている [8].

2.2 ΔBLEU: Discriminative BLEU

ΔBLEU [3] は，雑談応答生成のような出力多様性の高いテキスト生成のための半自動評価手法である．雑談応答生成では Twitter などのオンライン上の人の対話を評価に利用することが多いが，参照応答として利用できるのは基本的に特定の個人が行った一応答のみであるため，応答の多様性を考慮することが困難である．人手で参照応答として妥当な応答を書き尽くすことも現実的でないため，入力発話-参照応答ペアと発話と応答がそれぞれ類似する発話-応答ペアの応答を Twitter 上の大規模対話から疑似応答として収集することが行われている [14] が，人の評価との相関は依然，低い．そのため ΔBLEU では，疑似応答を収集して構築した参照応答に入力発話に対する応答としての妥当性を人手で付与し．BLEU 計算の重み付けとして利用することで，より人の評価との相関の高い評価尺度を実現している．

ΔBLEU では既存研究 [14] に倣って BM25 [11] を類似度関数に用いて，入力発話-参照応答に類似する発話-応答ペアを収集する．発話-応答ペアの類似度は入力発話と発話の類似度，参照応答と応答の類似度をそれぞれ計算して掛けることで計算される．疑似応答に対して付与する妥当性はクラウドーソーシングで収集したリッカート尺度風の 5 段階評価を $[-1, 1]$ の値に正規化して用いる．以上により入力発話 i に対して獲得した疑似応答とその妥当性 $w_{i,j}$ を利用して，式 (1) に用いる n-gram 精度 p_n を以下のように計算する．

$$\frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_{j: g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i)\}} \tag{4}$$

この式は式 (3) の各 n-gram g について，参照応答 h_j の妥当性で重み付けをした評価式となっている．

ΔBLEU では，参照応答の入力発話に対する応答としての妥当性を人手で付与するため，そのコストが問題となる．雑談対話はオープンドメインであるため，その評価は様々なドメインで行われるべきであるが，多様なドメインの発話に関する応答に人手で妥当性を付与することはコスト的に現実的でない．また，疑似応答の収集において，応答の類似性を考慮していること（さらにはその類似度計算に単語の一致に基づく BM25 を用いてること）から，内容的にも多様となりうる応答の多様性を考慮することが難しい．

3. 提案手法

本節では 2.2 節で述べた ΔBLEU の問題点を解決するために，収集した疑似応答候補に対して，Twitter 上に存在する複数応答を持つ発話を学習データの正例として用いた分類器により妥当性の自動付与を行う．これにより，大規模対話データを元に雑談応答を自動評価する手法を提案する．さらに，発話の類似性のみから疑似応答を収集することで応答の多様性を確保する．このような収集では妥当でない疑似応答が混入する可能性があるが，上記の分類器を流用することでフィルタリングし，疑似応答の質を担保する．

3.1 多様な疑似応答候補の収集

ΔBLEU では入力発話-参照応答に類似する発話-応答ペアの応答を疑似応答として収集した．しかし，発話に対する応答としては実際に行われた応答と内容が大きく異なる発話でも応答として成立しうる．このため疑似応答の収集において，参照応答との（表層的）類似性を考慮してしまうと，内容的に多様な応答候補を収集しにくくなってしまう．

そこで本研究では，入力発話のみを手がかりとし，入力発話と類似する発話に対する任意の応答を疑似応答候補として収集する．発話の類似性のみを手がかりとして疑似応答の収集を行うと，応答として不適切な疑似応答が混入する可能性が高まるが，この点については，3.2 節で述べる応答の妥当性を評価する分類器を流用してフィルタリングすることで解決する．

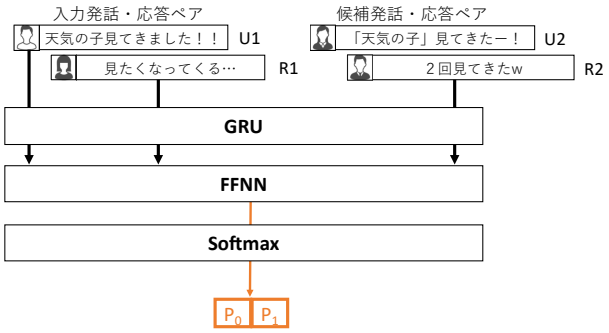

図 1　疑似応答の妥当性判定を行う分類器
Fig. 1　A classifier for judging the appropriateness of pseudo responses.

発話の類似性の判定についても，BM25 より柔軟に入力発話と内容の類似した発話を収集するために，分散表現ベースの手法を用いることを提案する．具体的には，事前に発話のベクトル表現を計算してそのコサイン類似度を用いて入力発話と類似する発話（とその応答）を収集する．発話ベクトルは，発話を構成する単語（トークン）のベクトル表現を集約することにより計算する．本稿では，予備実験の結果をもとに，WordPiece [17] によってトークン化を行い，事前学習された BERT [2] を用いて単語ベクトル化を行ったのち，単語ベクトルを平均することで発話ベクトルを得た．

3.2　分類期に基づく疑似応答候補の選別と妥当性評価

3.1 節で収集した疑似応答候補は，応答元となる発話の類似性のみに基づいて収集されるため，入力発話に対する応答としては不適切なものが含まれる可能性がある．また，ΔBLEU の適用のためには，疑似応答には入力発話に対する応答としての妥当性が評価されている必要がある．そこで，与えられた入力発話-参照応答ペアに対し，収集した各疑似応答候補の入力発話に対する応答としての妥当性を，教師あり学習に基づく分類器により評価する. 具体的に分類器は，入力発話-参照応答ペア，および収集した発話-応答ペアを入力して，発話-応答ペアの応答が入力発話に疑似応答とんなりうる確率値を計算し，[-1,1] に正規化して出力する．

この際，疑似応答の選別・妥当性評価を行う分類器の学習に用いる学習データをどのように得るかが問題となる．本研究では，疑似応答の収集に用いる Twitter の大規模性を最大限活用し，応答を複数持つ発話に着目して学習データ（正例）の収集を行う．具体的には，複数の応答を持つ発話について，一つの応答を参照応答，それ以外の応答を擬似応答候補とみなして発話が共通した 2 つの発話-応答ペアを生成し，分類器の正例として収集する．なお，負例についてはランダムに抽出した独立な発話-応答ペアを用いる．

分類器はニューラルネットワークを用いて学習する（**図 1**）．具体的には，入力発話 U1・参照応答 R1 と疑似応答 R2 (応答元の発話 U2) から U1・R1・R2 および U2・R2・R1 の組み合わせで結合して，正例または負例を 2 例を作る．次に，各例を Gated Recurrent Unit (GRU) [1] により 3 つ組のベクトルへと変換する．最終的に，それらを Feed-Forward Neural Network (FFNN) に入力し，その出力をソフトマックス関数に入力して U1 または U2 に対して R1 と R2 が交換可能となる（言い換えると正例および負例となる）確率をそれぞれ出力する．本モデルの学習時の損失は，FFNN の出力である正解ラベルへの確率との誤差によりそれぞれ計算する．

実際に入力発話 U1・参照応答 R1 と疑似応答 R2 (応答元の発話 U2) に対する最終的な評価値を得る際には，(U1, R1, R2) および (U2, R1, R2) に対する出力結果を各確率ごとに max を取って大きい方を出力する（負例に関しては，[-1,1] への正規化のために -1 を掛けて出力する）．このような計算をするのは，訓練データの正例で U1 と U2 が同一であることから，U1 と U2 を同時に考慮して学習すると過学習が起きる可能性が高いためである．

4.　実験

本節では，Twitter 上の大規模英語対話データセットを用いて提案評価手法の評価を行う．

4.1　大規模英語対話データセット

実験で利用する大規模英語対話データセットは，著者らの研究室で Twitter API[*1] を利用して 2011 年 3 月から継続的に収集している多言語 Twitter アーカイブから構築した．本アーカイブは著名な日本人ユーザ 30 名程度を選択し，それらがメンションもしくはリツイートしたユーザをさらに収集対象に追加することでユーザ数を順次拡大するとともに，その投稿を定期収集したデータである．

まず多言語 Twitter アーカイブから英語の投稿を選択する．収集された投稿には Twitter API が提供する言語判定結果が付与されているが，言語判定の信頼性を高めるため，これとは別に Twitter に特化した言語判定モデル ldig[*2] による言語判別も行った．ldig で提供されている Twitter 用のモデルでは 19 言語に対して 99%の分類精度で言語判定が可能である．Twitter アーカイブ中の投稿から，両言語判定結果で英語として判定された投稿のみを利用した．

次にこのようにして得られた英語投稿から，メンションもしくはリツイート以外の投稿を発話，それに対するメンションを応答とした発話-応答ペアを抽出し，英語対話データセットを構築した．一つの発話に複数の応答が存在する場合，各応答とのペアを一対話として抽出するが，応答が 4 つ以上存在する発話は評価する雑談応答生成モデルや疑似応答の分類器の訓練時に問題となる可能性があるためデー

[*1] https://dev.twitter.com/overview/api
[*2] https://github.com/shuyo/ldig

表 1　ハイパーパラメータ設定
Table 1　Hyparameter Settings.

雑談応答モデル	学習率	10^{-5}
	最適化方式	Adam
		$\beta_1 = 0.9$
		$\beta_2 = 0.98$
	学習率減衰	逆平方根
疑似応答分類器	埋め込み層・	512 次元
	隠れ層の次元 (GRU)	1024 次元
	隠れ層の層数 (FFNN)	5
	ドロップアウト	0.2
	学習率	0.001
	最適化方式	Adam [4]
		β_1 : 0.9
		β_2 : 0.999
	損失関数	交差エントロピー
	エポック数	15
	バッチサイズ	1000

タセットから削除した．

最終的に，構築された対話データは応答生成モデルの学習のために SentencePiece [5] を用いてトークン化を行った．このようにして得られた対話データセットの中から，本研究では 2018 年内に投稿された英語による発話-応答ペア約 5000 万対を雑談応答生成モデルの学習・評価，評価のための疑似応答（候補）の収集，疑似応答の妥当性評価に用いる分類器の学習・開発に用いた．以下で順に述べる．

4.2　雑談応答生成モデル

提案評価手法の評価のため，評価対象となる雑談応答生成モデルの学習を行った．具体的には，Pytorch[*3]で実装されたテキスト生成ライブラリ fairseq[*4]の Transformer [16] を使用して雑談応答モデルの訓練を行った．このモデルの学習データには 2018 年 1 月中に投稿された対話データから 200 万対の発話-応答ペアを，開発データには 1 万対の発話-応答ペアをランダムに選んで利用した．Transformer のハイパーパラメタは**表 1** に示されるもの以外は元文献 [16] と同一の値に設定した．

雑談応答モデル（および自動評価尺度の）の評価データとしては，100 の発話-応答ペアをランダムに選んだ．また，Trandformer による生成応答について，著者のうち一人によって既存手法 [3] に倣って 5 段階評価を行い，[-1,1] の範囲に正規化した．

4.3　応答妥当性分類器の学習

次に，応答の妥当性評価のための分類器の学習を行った．分類器の学習データとして 1000 万対，開発データとして 1 万対の発話-応答ペアのペア (正例・負例は同数) をランダムに選び，疑似応答を収集する対話データとして用いた．表 1 に学習で利用したハイパーパラメタを示す．分類器のパラメタは開発データで最小の損失を得たものを利用した．

4.4　比較手法

本研究では，ΔBLEU に対して，疑似応答の収集・選別方法および妥当性評価方法を新たに提案しているため，これらの要素をそれぞれ変えて提案手法と BLEU および ΔBLEU の比較を行う．

まず，発話-応答ペアの類似性を BM25 で計算する既存収集手法 [14] と発話のみの類似性を分散表現ベースの手法で計算する提案収集手法で疑似応答をそれぞれ収集し，得られた疑似応答を BLEU での評価に直接用いて，どちらの手法が疑似応答として妥当か評価を行う．ここで BLEU を用いる手法は，疑似応答をそのまま参照として用いて BLEU を計算する [14] と同じものである．

次に，BLEU (ΔBLEU で擬似応答の重みを全て 1 とみなしたもの)，ΔBLEU（疑似応答を人手評価），提案手法 ΔBLEU-auto（疑似応答を分類器で評価）について，疑似応答を（人手または分類器による）妥当性評価値でフィルタリングする場合とそうでない場合，それぞれについて人手評価との相関を比較する．全ての手法で修正 n-gram 精度の計算は，既存手法 [3] に倣って $n \geq 2$ (BLEU-2) を用いた．

なお，人手または分類器の妥当性評価値を用いて擬似応答のフィルタリングを行う場合は，0 以上の評価値が付与された疑似応答のみを利用する．また BLEU については人手および分類器の妥当性評価それぞれでフィルタリングした疑似応答を用いた場合の結果を示す．人手評価との順位相関の計算は，ΔBLEU にに倣って Kendall の τ および Spearman の ρ を利用した．

4.5　評価手順

4.2 節で述べた評価データに対する疑似応答候補は，分類器の学習・開発データを除いた対話データセット中で複数応答のない発話-応答ペアから収集を行った．まず，既存研究 [3] に倣い，各入力発話について対話データセットから発話が類似する上位 15 対の発話-応答ペアを取り出し，その応答を疑似応答候補として収集した．次に抽出した疑似応答を訓練済み分類器を用いて分類し，入力応答に対する応答としての妥当性を評価値として付与した．ただし，ΔBLEU とは異なり，入力発話自体を疑似応答に追加して用いることは行わず，参照応答は妥当性の評価値を 1 として利用した．結果として，最終的な参照応答としては元の参照応答に収集した擬似応答を加えた計 16 対を最大で利用した．なお，評価手法の評価の際には，Stanford NLP Tokenizer[*5]によりトークン化を行った．

[*3] https://pytorch.org/
[*4] https://ai.facebook.com/tools/fairseq/
[*5] https://github.com/stanfordnlp/stanfordnlp

表 2　雑談応答モデルに対する，参照応答とする疑似応答の収集方法を変えた BLEU による評価と人手評価との相関

Table 2　Rank correlation between BLEU and human judgements while varying the method of collecting pseudo responses.

評価手法	参照応答	(収集手法)	Kendall's τ	(p-value)	Speaman's ρ	(p-value)
BLEU	single		−0.012	(0.88)	−0.019	(0.85)
BLEU	all	(BM25)	0.110	(0.14)	0.147	(0.14)
BLEU	all	(提案手法)	**0.227**	(< 0.01)	**0.297**	(< 0.01)

表 3　雑談応答モデルに対する，BLEU, ΔBLEU, ΔBLEU-auto による評価と人手評価との相関（括弧内に p 値を示す）

Table 3　Rank correlation between each evaluation metrics and human judgements with or without filtering of pseudo responses (The p-value is shown in the parenthesis).

評価手法	参照応答	Kendall's τ	(p-value)	Speaman's ρ	(p-value)
BLEU	single	−0.012	(0.88)	−0.019	(0.85)
BLEU	$w \geq 0$	0.160	(0.04)	0.208	(0.04)
BLEU	$\hat{w} \geq 0$	0.284	(< 0.01)	0.375	(< 0.01)
BLEU	all	0.227	(< 0.01)	0.297	(< 0.01)
ΔBLEU	$w \geq 0$	0.143	(0.06)	0.190	(0.06)
ΔBLEU	all	0.022	(0.77)	0.026	(0.80)
ΔBLEU-auto	$\hat{w} \geq 0$	**0.288**	(< 0.01)	**0.383**	(< 0.01)
ΔBLEU-auto	all	0.178	(0.02)	0.232	(0.02)

4.6　結果

表 2 に，BLEU による評価で参照応答とする疑似応答の収集方法のみを変えた場合の結果を示す．なお single は元の参照応答のみを用いた結果である．結果から，疑似応答なしでは BLEU が評価尺度としてまともに動作しないことと，疑似応答の収集方法として提案手法が BM25 より適していることが確認できる．このため，次に述べる ΔBLEU と提案手法での比較では分散表現ベースの提案収集手法で疑似応答（候補）を収集し，比較を行う．

次に，**表 3** に BLEU，ΔBLEU，提案手法 ΔBLEU-auto の人手評価との相関を示す．$w \geq 0$ または $\hat{w} \geq 0$ は，人手による妥当性評価値 w と分類器による妥当性評価値 $\hat{w}$ でそれぞれ疑似応答のフィルタリングを行った場合の結果である．分類器により疑似応答のフィルタリングを行た提案評価手法により最も高い人手評価との相関が得られた．また分類器により正例として推定された疑似応答のみを参照応答として利用した BLEU も同等の相関を得た．一方で分類器により評価した疑似応答を全て利用した場合や，ΔBLEU は BLEU で得られた相関よりも低い相関となっている．

表 3 においての BLEU と人手評価との相関から，人手による妥当性の評価付与と提案手法による分類器による評価付与の違いを考察する．我々が提案する疑似候補収集手法では，分散表現ベースで発話のみの類似により疑似応答を収集するため単純に応答として見た場合は不適当な疑似応答が含まれる可能性がある．このため人手付与される評価値が低くなり，フィルタリングの結果残った疑似応答文が少なかったのではないかとか考えられる．実際，人手評価付与の場合フィルタリング後に平均 9.2 文の疑似応答が利用されたのに対し，分類器による自動評価付与では平均 13.7 文の疑似応答が利用されている．一方で，分類器による評価付与では，多少の文脈的不一致があったとしても，部分的な文字列における疑似応答としての有用性を判断して，人手評価との相関が低くても評価に有用な疑似応答に高い評価値を付与し，活用できたのではないかと考えられる．

5.　関連研究

本章では，雑談対話でも用いられる自動要約システムの自動評価尺度 ROUGE [7] と，参照応答とシステム出力に対する人手評価から評価関数を学習する評価手法 Adem [9] と RUBER [15] について紹介する．

ROUGE [7] は表層類似性に基づく自動要約システムの標準的な自動評価尺度である．ROUGE では BLEU を元に開発された評価尺度であるが，自動要約という問題の性質を考慮して，システム出力に含まれる n-gram の精度ではなく，参照応答の n-gram の再現率を評価するように変更したものである．ROUGE を雑談対話に利用した際，BLEU と同様に雑談の応答の多様性を考慮できないために，人手評価との相関が低いという問題が指摘されている [8].

次に，Adem [9] は人手評価を学習データとして用いた雑談生成応答のための評価関数学習手法である．この手法ではあらかじめ用意した人手評価に類似するようにニューラルネットワークによる学習を行う．この手法では，入力発話，参照応答，システム出力（と人手評価）を利用して，

評価関数を学習する．この手法では，少数の人手評価を学習データとして用いるため一定のコストがかかるほか，評価器が学習データのドメインに対して過学習する可能性がある．

RUBER [15] は参照応答を用いた評価手法と用いない評価手法を組み合わせた自動評価手法である．参照応答を用いる評価手法では参照応答と生成応答のベクトル表現における類似度を用いる．参照応答を用いない評価手法では，ニューラルネットワークを用いて負例サンプリングを用いた学習により入力発話に対する生成応答の妥当性を評価する．本手法は，応答多様性を明示的に考慮した手法でないため，本研究で行った予備実験では比較対象としなかった．詳細な比較は今後の課題とする．

6. おわりに

本稿では大規模な Twitter データを利用し，発話の類似性のみに基づいて疑似応答候補を収集し，これを自動獲得した学習データを用いて学習した分類器によって妥当性評価と選別することで，ΔBLEU を自動化する手法を提案した．Twitter から構築した大規模対話データセットを用いた実験を通して，人手によるアノテーションである ΔBLEU 以上の人手評価との相関が達成できることを確認した．

今回行った実験は 100 対の英語の発話-応答ペアに対し，著者一名による人手評価を用いた予備的な段階に留まっている．今後の課題として，より複数の言語について，より大規模な評価データを構築して，信頼性の高い評価を行うことを検討している．

謝辞　本研究の一部は，JST，CREST，JP-MJCR19A4 の支援を受けたものです．また，この研究の一部は 2019 年度国立情報学研究所 CRIS 委託研究の助成を受けています．

参考文献

[1] Cho, K., van Merriënboer, B., Bahdanau, D. and Bengio, Y.: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111.

[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186 (2019).

[3] Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J. and Dolan, B.: deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Short Papers*, pp. 445–450 (2015).

[4] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the third International Conference on Learning Representations (ICLR)* (2015).

[5] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pp. 66–71.

[6] Li, J., Galley, M., Brockett, C., Gao, J. and Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 110–119 (2016).

[7] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Text Summarization Branches Out*, pp. 74–81 (2004).

[8] Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y. and Pineau, J.: Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1116–1126 (2017).

[9] Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y. and Pineau, J.: Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses, pp. 1116–1126 (2017).

[10] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation, *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318.

[11] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M. and Gatford, M.: Okapi at TREC-3, *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126 (1995).

[12] Sato, S., Yoshinaga, N., Toyoda, M. and Kitsuregawa, M.: Modeling Situations in Neural Chat Bots, *Proceedings of ACL 2017, Student Research Workshop*, pp. 120–127 (2017).

[13] Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. and Bengio, Y.: A Hierarchical Latent Variable Encoder-decoder Model for Generating Dialogues, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3295–3301 (2017).

[14] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J. and Dolan, B.: A Neural Network Approach to Context-Sensitive Generation of Conversational Responses, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 196–205 (2015).

[15] Tao, C., Mou, L., Zhao, D. and Yan, R.: RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pp. 722–729.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems (NIPS) 30*, pp. 5998–6008.

[17] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. and Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, *CoRR*, Vol. abs/1609.08144 (2016).