The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

自動拡張した参照応答に基づく雑談対話システムの自動評価

Automatic evaluation of open-domain dialogue systems using automatically-augmented references

蔦 侑磨 *1 Yuma Tsuta　吉永 直樹 *2 Naoki Yoshinaga　豊田 正史 *2 Masashi Toyoda

*1東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

*2東京大学生産技術研究所
Institute of Industrial Science, the University of Tokyo

In open-domain dialogues, the content and style of responses can vary. However, it is difficult to consider the diversity of responses when evaluating responses generated by dialogue systems, since basically only one response can be extracted as a reference response from real conversations. To address this problem, ΔBLEU uses reference responses that are extended with responses in massive dialogue data and are manually annotated with appropriateness as a response. Because the human annotation is costly, we cannot utilize ΔBLEU for a large-scale evaluation of open-domain dialogue systems that should be evaluated in various contexts. We propose a fully-automatic evaluation method ΔBLEU-auto that annotates the appropriateness of extended responses used in ΔBLEU by a classifier trained with automatically-collected training data. Experimental results confirmed that ΔBLEU-auto is comparable to ΔBLEU in terms of correlation with human judgement, and also improves the state-of-the-art evaluation method, RUBER, by integrating our ΔBLEU-auto into RUBER.

1. はじめに

Apple Siri や Amazon Alexa, Google Assistant, LINE Clova など人と会話を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 [Li 16, Serban 17, Sato 17, Liu 19] が盛んに行われるようになった．

雑談対話システムで中心的に研究される雑談応答生成タスクでの問題点として，生成応答に対する自動評価手法が確立していないことが挙げられる．雑談応答生成タスクの評価で用いられる BLEU [Papineni 02] や ROUGE [Lin 04] などの自動評価手法は元々，機械翻訳や自動要約などの雑談対話とは別のテキスト生成タスクに対して設計されたものである．これらを雑談応答生成モデルの評価に用いた場合，人手評価との相関が低くなることが問題として指摘されている [Liu 16]．これは，機械翻訳や自動要約に比べて雑談対話ではスタイル・内容共に多様な出力 (応答) が許容されるにも関わらず，実会話を評価用対話データとして用いた場合，基本的に一応答のみしか参照応答として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半自動評価手法として ΔBLEU [Galley 15] が提案されている．この手法ではまず，Sordoni らの参照応答の拡張手法 [Sordoni 15] に倣い，Twitter 上の大規模対話データセットから類似会話の応答を参照応答に追加して利用することで応答多様性を考慮する．次にこの拡張参照応答に応答としての妥当性を人手で付与することで拡張参照応答の品質を考慮した評価を行う．ΔBLEU での問題点は人手評価による妥当性付与のコストであり，これをオープンドメインな雑談応答生成タスクの評価に足る大規模評価用対話データに行うことは現実的でない．また，近年提案されている雑談応答生成タスクの自動評価手法 [Lowe 17, Tao 18, Ghazarian 19] では BLEU や ΔBLEU と同様の着眼点である参照応答と生成応答の比較による評価以外に，発話と応答の関連性による評価を考慮し，モデル化している．しかし，ΔBLEU と同様に応答の多様性にまで明示的に考慮した手法は存在しない．

本研究では ΔBLEU の問題を解決するために，Twitter を利用して自動収集した拡張参照応答に対し，自動生成した教師ラベル付き対話データで学習した分類器によって応答妥当性の自動付与を可能にする評価手法 ΔBLEU-auto を提案する．ΔBLEU では，入力発話-参照応答ペアと発話・応答の両方が類似する発話-応答ペアの応答を拡張参照応答として収集しているが，応答の類似性まで考慮した収集を行うと，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，拡張参照応答の収集の際に発話のみの類似性に基づき応答を収集することで，より多様な応答を収集することを試みる．さらに提案手法では，既存の自動評価手法 RUBER [Tao 18] でモデル化される発話と応答の関連性による評価を考慮していないため，RUBER に提案手法を組み込んだ自動評価手法の有効性を確認する．

実験では Twitter 上の日本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，拡張参照応答の獲得，および拡張参照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，提案評価手法の評価を行った．

連絡先: e-mail アドレス tsuta@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 [Galley 15], 並びに ΔBLEU の基礎となる BLEU [Papineni 02] について説明する．

2.1 BLEU

BLEU [Papineni 02] は機械翻訳タスクにおける標準的な自動評価手法である．BLEU は，システム出力と参照出力で重複する n-gram の出現回数を用いた表層類似性に基づく評価を行う．具体的に評価値は，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．

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

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

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

ここで η, ρ はそれぞれシステム出力と参照出力の平均文長，n, N は n-gram の n とその任意の最大値，$r_{i,j}$ は各 i 番目の入力に対する j 番目の参照出力，h_i はそのシステム出力，$\#_g(u)$ は文 u における n-gram g の出現回数，$\#_g(u,v)$ は $\min\{\#_g(u), \#_g(u)\}$ を意味する．

BLEU を雑談応答生成タスクの評価に用いた場合，人手評価との相関が低いことが指摘されている [Liu 16]．これは，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，実会話を利用した評価では，基本的に一応答のみしか参照応答に利用できないなどの要因が考えられる．

2.2 ΔBLEU: Discriminative BLEU

ΔBLEU [Galley 15] は，雑談応答生成タスクのような出力多様性の高いテキスト生成タスクのための半自動評価手法である．雑談応答生成タスクでの実験として，応答の多様性を考慮した評価のために，大規模対話ログを利用して参照応答を拡張し，入力発話に対する拡張参照応答の妥当性を人手により評価して，これらを生成応答の評価に利用する．

具体的に，ΔBLEU では既存研究 [Sordoni 15] に倣って BM25 [Robertson 94] 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発話同士，また応答同士の類似度をそれぞれ計算して掛け合わせることで計算される．この収集した応答と，入力発話，参照応答を拡張参照応答とし，さらに入力発話に対する応答としての妥当性を人手により付与する．なお，拡張参照応答に対して付与する妥当性は 5 段階のリッカート尺度を $[-1, 1]$ の値に正規化して用いる．以上により入力発話 i に対して獲得した拡張参照応答 $r_{i,j}$ 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．

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

この式は式 (3) の各 n-gram g について，拡張参照応答 $r_{i,j}$ の妥当性 $w_{i,j}$ で重み付けをした評価式となっている．

ΔBLEU では，拡張参照応答の事例ごとに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，雑談応答生成タスクの評価は様々なドメインで行われるべきである．しかし，多様なドメインの発話に対する応答に，網羅的に人手で妥当性を付与することは現実的でない．また，参照応答拡張のための類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 を用いていることから，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と考えられる．

3. 関連研究

本章では雑談応答生成タスクでの自動評価手法である，生成応答に対する人手評価を教師データとして評価関数を学習する ADEM [Lowe 17] と，発話と応答の関連性評価を教師データなしに学習可能にした RUBER [Tao 18] について説明する．

ADEM [Lowe 17] では，入力発話，参照応答，システム出力を入力として，人手評価を模倣するように評価関数を学習する．ADEM の問題点として，人手評価のコストがかかることや，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ことが問題点として挙げられる．

RUBER [Tao 18] は二つの評価手法を組み合わせた自動評価手法であり，ADEM と異なり学習時に人手評価を必要としない．評価手法の一つは参照応答と生成応答の類似性による評価を行い，これらのベクトル表現のコサイン類似度により評価値を計算する (Referenced Scorer)．もう一つの評価手法では入力発話と生成応答の関連性による評価を，負例サンプリングによる教師なし学習モデルにより推定する (Unreferenced Scorer)．この二つの評価値を組み合わせることで総合的な評価値を算出する．Unreferenced Scorer の改善方法として，BERT [Devlin 19] による単語のベクトル表現を事前学習済みベクトルとして利用することで，RUBER の人手評価との相関が向上することが確認されている [Ghazarian 19]．Referenced Scorer は，BLEU や ΔBLEU と同じ参照応答との比較を目的とした評価手法であるため，これらと置換が可能である．本研究では，RUBER と同じ自動評価手法である提案手法との組み合わせを検討した．

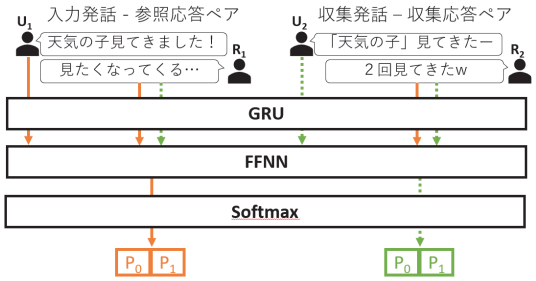

図 1: 収集応答の妥当性判定を行う分類器

4. 提案手法

本章では 2.2 節で述べた ΔBLEU の問題点を解決するために，ある会話（入力発話-参照応答ペア）の発話に対して別の会話 (収集した発話-応答ペア) の応答が利用できるかを判別する分類器を用いて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生成タスクでの自動評価手法を提案する．さらに，参照応答拡張のために発話の類似性のみに基づいて応答を収集することで応答内容の多様化を試みる．

4.1 参照応答拡張のための多様な応答の収集

ΔBLEU では参照応答拡張のために入力発話-参照応答に類似する発話-応答ペアの応答を収集した．しかし，実際に行われた参照応答と内容が大きく異なる応答でも入力発話に対する応答として成立しうる．このため，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を収集しにくくなってしまう．そこで本研究では，発話類似性のみに基づく応答の収集を試みる．また文間の類似度判定についても，BM25 より柔軟に意味的類似性を考慮するために，分散表現ベースでの類似度判定による収集を提案する．

具体的には，事前に文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．文ベクトルは，文を構成する単語（トークン）のベクトル表現を平均することにより計算した．ΔBLEU に倣い，以後，収集した応答と入力発話，参照応答を入力発話に対する拡張参照応答と呼ぶ．

4.2 分類器を利用した拡張参照応答への妥当性付与

提案手法では 4.1 節で収集した拡張参照応答に，ΔBLEU での人手による妥当性付与の代わりに，分類器を用いて発話に対して応答となりうる確率をの妥当性として自動付与する．具体的に分類器は，入力発話-参照応答ペア (U_1 ・R_1)，および収集した発話-応答ペア (U_2・R_2) を入力して，収集した応答 (R_2) が入力発話 (U_1) の応答となりうる（つまり U_1 - R_2 が成立する）確率を出力する．この確率を入力発話に対する応答の妥当性とするが，ΔBLEU に合わせるため，[-1,1] に正規化を行う．

この妥当性評価を行う分類器を訓練するために，学習用データが必要である．本研究では，複数の応答を持つ発話について，一つの応答を参照応答，それ以外の応答を収集した応答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，正例として収集する．そして，ランダムに抽出した 2 つの独立な発話-応答ペアを負例として用いる．

分類器（図 1）はニューラルネットワークを用いて学習する．訓練データの正例で入力する 2 文の発話が同一であることから，この特徴に過学習する可能性がある．そのため，2 文の発話が独立になるように，入力する必要がある．具体的な手順で

はまず，入力発話 U_1・参照応答 R_1・収集発話 U_2・収集応答 R_2 から U_1・R_1・R_2 および U_2・R_2・R_1 の組み合わせで 3 つ組を 2 つ作る．次に，Gated Recurrent Unit (GRU) [Cho 14] により 3 つ組をベクトルへと変換し，結合することで 2 つのベクトルを得る．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 U_1 または U_2 に対して R_1 と R_2 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-参照応答ペア (U_1, R_1) と収集発話-収集応答ペア (U_2, R_2) のペアに対する最終的な評価値を得る際には，(U_1, R_1, R_2) および (U_2, R_1, R_2) 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）．

5. 実験

本節では，大規模 Twitter データから抽出した日本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．具体的に，雑談応答選択モデルや雑談応答生成モデルが出力する応答や実応答（5.2 節）を利用して，各評価手法 (5.4 節) の性能比較を行う．性能比較の際には，各評価手法の人手評価との相関によりその性能を定量化し比較する (5.5 節)．

5.1 大規模日本語対話データセット

実験で利用する大規模日本語対話データセットは，著者らの研究室で 2011 年 3 月から継続的に収集している Twitter アーカイブから構築した．具体的には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-応答ペアを抽出した．これらは応答生成手法や妥当性の自動付与のための分類器の構築，参照応答拡張のための収集先のデータベースとして利用した．また各学習データがテストデータに影響しないように，応答生成手法の学習データには 2018 年内のデータを，各評価手法の性能比較にも利用するテストデータには 2019 年内のデータを，提案手法で利用する分類器の構築と参照応答拡張のための応答の収集は 2017 年内のデータを利用した．

5.2 雑談応答生成モデル

本節では提案評価手法の評価のため，評価対象となる（応答を出力する）雑談応答選択・生成モデル，及び実応答について説明する．実験では応答選択モデルとして発話の類似性に基づいて応答を抽出する方法 [Liu 16] を，応答生成モデルとして VHRED [Serban 17] を，理想的な応答として参照応答とは異なる実応答を利用した．但し，情報検索型では TF-IDF の代わりに，その改善手法であり ΔBLEU でも利用される BM25 を流用した．応答生成手法に利用する対話データは一定以上継続した対話を利用するため，発話 U_i に対する応答 $R_{i,t1}$ に対してさらに応答 $R_{i,t2}$ が行われ，かつ，U_i と $R_{i,t2}$ が同一アカウントによる投稿である会話のみを抽出した．さらに実応答を生成応答として利用するため，2 つ以上の応答を持つ対話データから一つの応答を参照応答としてランダムに選択した．結果，応答生成手法に利用する対話データは訓練データに 240 万対の発話-応答ペアを，開発データに 1 万対の発話-応答ペアを利用した．応答生成モデルでは SentencePiece [Kudo 18] を利用して，応答選択モデルでは MeCab 0.996 (ipadic 2.7.0, 以下同)[*1] を利用して分かち書きを行った．

5.3 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は 4.2 節で述べた発話-応答ペアのペアであり訓練データに 560 万対，開発データに 1 万対を利用した（正例・負例は同数）．分類器に入力するデータは事前に SentencePiece により分かち書きを行った．分類器は 512 次元の単語埋め込み層，5 層 1024 次元の隠れ層から構築され，学習率 0.001，最適化手法 Adam [Kingma 15]，損失関数は交差エントロピー，バッチサイズ 1000，エポック数 15 で学習した．また，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．

5.4 比較評価手法

本研究では，ΔBLEU に対して，収集方法の変更と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，並びに RUBER との組み合わせによる評価を行う．

まず，参照応答拡張のための応答の収集手法として，類似対象と計算方法の変更としてそれぞれ，発話と応答の両方の類似性から発話のみの類似性，BM25 による類似度計算から分散表現のコサイン類似度に変更している．全ての組み合わせを検討することで，各収集手法で得られた拡張参照応答に基づく BLEU による評価との人手評価の相関を比較し，提案手法による収集手法の有効性を確認する．

次に，BLEU (ΔBLEU で拡張参照応答の重みを全て 1 とみなしたもの)，ΔBLEU（拡張参照応答を人手評価），提案手法 ΔBLEU-auto（拡張参照応答を分類器で評価）について，人手評価との相関を比較する．

最後に提案手法 ΔBLEU-auto を RUBER での Unreferenced Scorer と組み合わせた自動評価手法について元の RUBER (Referenced Scorer と Unreferenced Scorer を組み合わせた手法) と比較を行う．この際の Unreferenced Scorer の学習データとして，5.3 節でのデータをそのまま利用した．

5.5 評価手順

まず自動評価手法の評価データとして，複数応答を持つ発話を 100 発話ランダムに選び，各発話に対して応答をランダムに一つ選んで 100 対の発話-応答ペアを得た．次に，各生成応答手法 (5.2 節) により生成された生成応答，及び参照応答以外の実応答に対して，既存手法 [Galley 15] に倣って著者らの所属する研究室の日本語を母語とする（著者らを除く）学生 6 人によって生成応答の入力発話に対する適切さについて 5 段階のリッカート尺度による評価を行った．

次に，既存研究 [Galley 15] に倣い，5.4 節で述べたそれぞれの手法で類似する上位 15 対の発話-応答ペアを取り出す．具体的に参照応答拡張のための応答は，2017 年内の対話データセット中の発話-応答ペア約 1600 万対から収集した．これらのデータは MeCab により分かち書きを行い，分散表現を利用した収集 (4.1 節) では，このデータに対して Glove [Pennington 14] により学習された単語ベクトルを利用した．

そして拡張参照応答に訓練済み分類器 (5.3 節) もしくは人手により，入力発話に対する応答としての妥当性を付与した．人手評価の際には，日本語を母語とする 4 人により ΔBLEU と同様に妥当性評価を行った．

最後に，ΔBLEU, ΔBLEU-auto では拡張参照応答とそれぞれの妥当性評価を利用して，BLEU では参照応答のみ，もしくは拡張参照応答を利用して，RUBER では参照応答のみを利用して 3 種 100 対の各生成応答に対して評価を行った．各評価手法の性能評価はこの 300 件の生成応答毎への人手評価との相関により評価する．人手評価との相関の計算は Spearman's ρ と Pearson's r を利用し，6 人による人手評価との相関をそれぞれ計測した 6 つの相関係数中の最大値と最小値を示す．この人手評価との相関の最大値と最小値の差により評価手法の安定性を，相関の最小値により評価手法の最低限の性能を見積もることができると考えられる．修正 n-gram 精度を利用する全ての手法での計算は，MeCab[*1] により文を分かち書きし，既存手法 [Galley 15] に倣って $n \geq 2$ (BLEU-2) を用いた．

5.6 結果

表 1 に，BLEU による評価で参照応答拡張のための応答の収集方法のみを変えた場合の結果を示す．Pearson's ρ では手法に依らず参照応答を複数にすることにより人手評価との相関が向上することが確認できた．次に Spearman's r では（発話を対象として分散表現で類似度を計算する）提案手法以外は単一参照の BLEU に劣るが，提案手法ではいずれの手法よりも相関が高いことを確認できた．結果として，提案手法が BLEU

*1 https://taku910.github.io/mecab/

類似度計算 対象	手法	Spearman's ρ max	Spearman's ρ min	Pearson's r max	Pearson's r min
(人手評価)		0.773	0.628	0.778	0.607
(単一の参照応答)		0.186	0.091	0.276	0.190
発話と応答	BM25	0.257	0.138	0.298	0.173
発話	BM25	0.265	0.136	0.332	0.178
発話と応答	分散表現	0.265	0.136	0.332	0.178
発話	分散表現	**0.333**	**0.181**	**0.366**	**0.209**

表 1: 参照応答拡張のための応答収集手法を変えた場合の BLEU による評価と人手評価との相関.

評価手法	参照応答拡張	Spearman's ρ max	Spearman's ρ min	Pearson's r max	Pearson's r min
(人手評価)		0.773	0.628	0.778	0.607
BLEU		0.186	0.091	0.276	0.190
BLEU	✓	0.333	0.186	0.366	0.209
ΔBLEU-auto	✓	0.330	0.281	**0.394**	**0.332**
ΔBLEU	✓	**0.366**	**0.300**	0.360	0.294

表 2: 各評価手法による評価と人手評価との相関.

評価手法	Spearman's ρ max	Spearman's ρ min	Pearson's r max	Pearson's r min
(人手評価)	0.773	0.628	0.778	0.607
Referenced Scorer	0.188	0.071	0.075	0.016
ΔBLEU-auto	0.330	0.281	0.360	0.294
Unreferenced Scorer	0.342	0.225	0.336	0.217
+Referenced Scorer	0.339	0.206	0.325	0.193
+ΔBLEU-auto	**0.435**	**0.323**	**0.450**	**0.338**

表 3: RUBER に ΔBLEU-auto を組み込んだ評価手法による評価と人手評価との相関.

の拡張参照応答として適していることが確認できた．このため，以後の実験での提案手法では提案手法による収集手法により評価を行った．

次に，表 2 に BLEU，ΔBLEU，提案手法 ΔBLEU-auto の人手評価との相関を示す．まず，提案手法による自動付与した妥当性を利用することで，利用しない場合 (BLEU) に比べてほとんどの人手評価との相関が向上することが確認できた．一方で，ΔBLEU との比較では，Spearman's ρ では ΔBLEU が，Pearson's r では ΔBLEU-auto がより高い相関を得た．このことから，評価手法として ΔBLEU-auto が ΔBLEU と同等程度の性能であると考えられる．

最後に，表 3 に，RUBER と組み合わせた際の ΔBLEU-auto の人手評価との相関を示す．結果として，ΔBLEU-auto を Unreferenced Scorer と組み合わせることで人手評価との相関が向上し RUBER より高い相関が得られることが確認できた．

6. おわりに

本稿では大規模な Twitter データを利用し，ΔBLEU を自動化する手法 ΔBLEU-auto を提案した．具体的に，発話のみの類似性により参照応答を拡張し，これを自動収集したデータによる教師あり学習をした分類器によって妥当性評価を行う．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 と同等の人手評価との相関が達成できることを確認した．また，ΔBLEU-auto を RUBER と組み合わせることでより高い相関が得られることを確認した．

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

[Cho 14] Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y.: On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, in *SSST* (2014)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in *NAACL-HLT* (2019)

[Galley 15] 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, in *ACL-IJCNLP* (2015)

[Ghazarian 19] Ghazarian, S., Wei, J., Galstyan, A., and Peng, N.: Better Automatic Evaluation of Open-Domain Dialogue Systems with Contextualized Embeddings, in *NeuralGen workshop* (2019)

[Kingma 15] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, in *ICLR* (2015)

[Kudo 18] Kudo, T. and Richardson, J.: SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing, in *EMNLP, demo session* (2018)

[Li 16] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B.: A Diversity-Promoting Objective Function for Neural Conversation Models, in *NAACL-HLT* (2016)

[Lin 04] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, in *ACL-04 Workshop on Text Summarization Branches Out* (2004)

[Liu 16] Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J.: How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, in *EMNLP* (2016)

[Liu 19] Liu, C., He, S., Liu, K., and Zhao, J.: Vocabulary Pyramid Network: Multi-Pass Encoding and Decoding with Multi-Level Vocabularies for Response Generation, in *ACL* (2019)

[Lowe 17] 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, in *ACL* (2017)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: Bleu: A Method for Automatic Evaluation of Machine Translation, in *ACL* (2002)

[Pennington 14] Pennington, J., Socher, R., and Manning, C.: Glove: Global Vectors for Word Representation, in *EMNLP* (2014)

[Robertson 94] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M.: Okapi at TREC-3, in *TREC* (1994)

[Sato 17] Sato, S., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M.: Modeling Situations in Neural Chat Bots, in *ACL-SRW* (2017)

[Serban 17] 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, in *AAAI* (2017)

[Sordoni 15] 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, in *NAACL-HLT* (2015)

[Tao 18] Tao, C., Mou, L., Zhao, D., and Yan, R.: RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems, in *AAAI* (2018)