修士論文

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

東京大学大学院 情報理工学系研究科
電子情報学専攻
48-186443　蔦　侑磨

指導教員　　吉永　直樹 准教授

2020 年 1 月 30 日 提出

本論文は東京大学大学院情報理工学系研究科に修士号授与の要件として提出した修士論文である.

要約

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，雑談対話応答システムの評価に実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半自動評価評価手法 ΔBLEU や，人手により参照応答を拡張し，既存の評価手法を複数の参照応等に基づく評価手法へ拡張する研究が存在する．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参照応答に基づく評価手法のための妥当性つき複数参照応答を獲得する手法を提案する．実験では大規模な Twitter データを利用して，提案手法により妥当性付き複数参照応答を獲得し，雑談対話応答システムの評価に関して提案手法により獲得した妥当性付き複数参照応答を利用した評価の有効性を確認した．その結果，単一参照応答のみに基づく評価手法や妥当性を利用しない複数参照応答に基づく評価手法と比較して，人手評価との相関が向上することを確認した．さらに，妥当性付き複数参照応答に基づく自動評価手法を，雑談対話システムの自動評価手法として人手評価と最も相関の高い自動評価手法である RUBER と組み合わせることで，既存の組み合わせによる評価と比較して，人手評価との相関が向上することを確認した．

目次

図目次

表目次

第 1 章

序論

1.1 雑談対話システムと評価手法

Apple Siri や Amazon Alexa, Google Assistant や 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 [1–3] 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 [4] や ROUGE [5] などの自動評価尺度は機械翻訳や自動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 [6]．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，Galley らが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 [7] を提案している．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 [8] 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．

同様に Gupta らが，人手により拡張した参照応答を利用した，既存の参照応答に基づく評価手法の複数参照応答に基づく評価手法への拡張を提案しているが，依然としてオープンドメインな大規模評価データの構築にかかるコストは解消できていない．

1.2　本研究の目的と貢献

本研究では応答多様性を考慮した大規模評価データの構築にかかるコストを低減させるために，ΔBLEU での手法を参考に，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により妥当性付きの複数参照応答を獲得する手法を提案する．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 アーカイブ上の対話データを利用して，評価対象と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，提案評価手法の有効性を確認した．まず予備実験として，ΔBLEU と同様の実験を提案手法による参照応答の拡張方法，並びにそれらへの妥当性付与手法を比較して有効性を確認した．次に，Gupta らの参照応答に基づく評価手法の，複数参照応答に基づく評価手法への拡張を参考にして，BLEU 以外の単一参照応答に基づく評価手法へも適用し，より汎用的に提案手法の有効性を確認した．結果として単一参照応答のみに基づく評価手法や妥当性を利用しない複数参照応答に基づく評価手法と比較して，提案手法による妥当性付き複数参照応答を活用した評価手法が人手評価との相関が向上することを確認した．さらに，妥当性付き複数参照応答を利用した評価手法を，雑談対話システムの自動評価手法として人手評価と最も相関の高い自動評価手法である RUBER [9] と組み合わせることで，既存の組み合わせによる評価と比較して，人手評価との相関が向上することを確認した．

1.3　本論文の構成

以降の本論文の構成は以下の通りである．

第 2 章 本研究に関連する雑談対話システム並びに，雑談対話システムの評価に用いられる他の自然言語処理タスクから転用された自動評価手法について述べる．

第 3 章 本研究と関連する雑談対話タスクでの評価手法について述べる．

第 4 章 本研究の提案手法について述べる．

第 5 章 本研究での先行研究である ΔBLEU と比較した予備実験，並びに評価手法に汎用な妥当性付き複数参照応答としての有効性を確認した実験について述べる．

第 6 章 全体のまとめについて述べる．

第 7 章 今後の課題について述べる．

第 2 章

基礎知識

本章では本論文での実験で利用する雑談対話応答システム，並びに，生成モデルの評価に一般的に利用される評価尺度について説明する．

2.1 雑談対話応答システム

本節では本論文での実験で利用する雑談対話応答システムについて説明する．一般的な雑談対話システムには，クエリとなる発話に適切な応答をデータベース中から抽出し，この応答をそのまま出力する情報検索型モデルと，会話データから発話に対する応答を学習したモデルを利用して発話に対する応答を逐次的に生成する生成型モデルの二種類が存在する．本論文では実験過程で利用する雑談対話応答システムとして，情報検索型対話モデルのための類似度計算手法として Robertson らの BM25 [10] と，生成型モデルとして Serban らの VHRED [11] と Vaswani らの Transformer [12] を採用した．本節では，まず情報検索型モデルについて説明を行い，次に生成型モデルの VHRED について説明する．

2.1.1 情報検索型対話モデル

情報検索型の雑談対話モデルは，データベースとなる大規模対話コーパス中から入力発話をクエリとして，応答を検索して出力する手法である．出力する応答は実際に個人の行った実応答なので，文の流暢性を気にする必要がないことがこの手法の有用性である．一方で，入力発話に対する選択応答の関連性が保証されないことが問題としてあげられる．この情報検索型モデルとして TF-IDF 等の手法により発話文の類似性

評価を行う手法 [13] や発話-応答関連性を学習することで直接応答を選択する手法 [13] が存在する．今回の実験では TF-IDF の派生手法である BM25 を利用した，発話の類似性から応答を選択する手法を行ったため，BM25 について説明を行う．

BM25 [10] は文書検索分野において，データベース中の文書 D が与えられた際の単語列 $W = \{w_1, ..., w_n\}$ の重要度の算出に利用される．BM25 は Term Frequency (TF) と Inverse Document Frequency (IDF) を基に計算を行う．TF は文書 D 中での単語 w_i の出現頻度を，IDF は単語 w_i の文書集合中での特定の文書への偏在性を示している．TF と IDF を組み合わせた手法である TF-IDF とは異なり，各文書の長さを考慮した手法となっている．文書 D が与えられた際の単語列 $W\{w_1, ..., w_n\}$ の BM25 によるスコアは以下の式により求められる．

$$\mathrm{score}(D, W) = \sum_{i=1}^{n} \mathrm{IDF}(w_i) \cdot \frac{f(w_i, D) \cdot (k_1 + 1)}{f(w_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\mathrm{avgdl}})} \tag{1}$$

$$\mathrm{IDF}(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5} \tag{2}$$

$f(w_i, D)$ は文書 D における単語 w_i の出現頻度，$|D|$ は文書 D の単語数，N はデータベース中の全文書数，avgdl は文書の平均単語数，$n(w_i)$ は単語 w_i を含む文書数を表す．k_1, b は任意のパラメータである．

雑談対話システムでの評価時の利用として，Sordoni ら [14] は会話集合からテストデータの発話並びに応答の両方の BM25 スコアが高い発話-応答ペアを候補会話データから収集し，この応答を参照応答に加えて，拡張することで，複数参照を用いて雑談対話システムを評価した．具体的には τ をテストデータの会話データ，$\tilde{\tau}$ を候補会話データ，m を発話，r を応答，$d(i, j)$ を BM25 による文 j に対する文 i の重要度として以下の式のように計算し，会話 j に対する会話 i の類似度 $s(i, j)$ を求める．

$$s(\tilde{\tau}, \tau) = d(m_{\tilde{\tau}}, m_{\tau}) \left(\alpha d(r_{\tilde{\tau}}, r_{\tau}) + (1 - \alpha)\epsilon\right) \tag{3}$$

α は 0 以上 1 以下のパラメタ，ϵ は正のパラメタである．

2.1.2　VHRED: Variational Hierarchical Recurrent Encoder Decoder

近年一般的に利用される深層学習による文の生成モデルは再帰型ニューラルネットワーク (Recurrent Neural Networ; RNN) に基づいたモデルである．RNN に基づいた生成型モデルの雑談対話応答システムには，“It's OK” や “I don't know” といった当たり障りのない応答を行いやすいといった特徴がある．この特徴は，長い会話のやり取りを目的とした雑談対話タスクの応答にはふさわしくないとして問題視されている [1]．Serban らの提案した VHRED [11] は先行研究である会話の文脈を考慮した生成型モデル HRED [2] での普遍的な応答の生成という同様の問題点を，モデルの出力機構に対して確率的なノイズを付与することで解決した．

VHRED の説明にあたってまず，文の生成モデルで一般的な RNN による逐次的な文字生成の仕組みについて説明する．RNN は文字の入力または出力，あるいはその両方を同時に逐次的に行える機構であり，モデルパラメタ θ の RNN は文字列 $W\{w_1, ...w_M\}$ の確率分布を以下の式のように解釈することでモデル化を行う．

$$P_\theta\left(w_1, ...w_M\right) = \prod_{m=2}^{M} \quad P_\theta\left(w_m | w_1, ..., w_{m-1}\right) P_\theta\left(w_1\right) \tag{4}$$

RNN は各タイムステップでの再帰的な処理を行い，ある時刻 t における入力 w_t に対する隠れ層 h_t を以下の式により出力する．

$$h_t = f_\theta\left(h_{t-1}, w_t\right) \tag{5}$$

ここでの f_θ は非線形関数を示し，モデルによって活性化関数 (activation fnction) やゲート関数 (gating function) と呼ばれる．これにより，隠れ層 h_t を時刻 t 以前のトークン列が集約されたパラメタとして表し，これを利用して次に出力するトークン w_{n+1} を以下の式による確率分布として表現する．

$$P_\theta\left(w_{t+1} | w_1, ...w_t\right) = P_\theta\left(w_{t+1} | h_t\right) \tag{6}$$

出力される単語は特定の単語集合 V 中から選択されるものであると仮定して，出力 w_{t+1} の確率分布は主にソフトマックス関数に対して h_t を入力することで求める．この RNN の拡張として隠れ層での入出力情報を調整可能にした Long Short-Term Memory (LSTM) [15] や LSTM をの構造を簡素化した Geted Recurrent Unit

(GRU) [16]，順方向・逆方向の時系列の双方向での計算から求められるパラメタを利用する Bidirectional RNN (Bi-RNN) [17] が存在する

次に RNN を生成型モデルとして用いる場合，発話のトークン列 $X = \{x_1, ..., x_N\}$ を入力データとして，その発話に対する応答のトークン列 $Y = \{y_1, ..., y_M\}$ を出力データとして，これを復元できるように学習を行う．これを実現する単純な生成型モデルの構造として，発話列 X を受け取るエンコーダとしての RNN と応答列 Y を出力するデコーダとしての RNN の 2 つの RNN により構築されるモデルがあげられる．この際，モデルのパラメータは応答列 Y のパープレキシティ (Perplexity; PPL) 等を目的関数としてパラメータを学習する．

$$\mathrm{PPL}\left(w_1, ..., w_M\right) = \prod_{i=1}^{M} P_\theta(w_i)^{-\frac{1}{M}} \tag{7}$$

雑談対話では複数回のやり取りとなる長い会話の文脈を考慮する必要があるが，2 文間のみを対象とする単純な手法ではこれが考慮できない．Serban らはこれを可能にする手法 HRED [2] を提案した．この手法では 2 段階の階層的な RNN の構造を有している．一つは上記と同様のトークン列としての文を扱う RNN であり，もう一つは文の系列としての会話履歴を扱う RNN(contextual-RNN) である．後者の RNN は前者の RNN の最終隠れ層を入力として，各文の履歴を集約する RNN となっている．これにより，文生成時に過去の発話履歴を加味した応答が可能になっている．具体的には以下のような式をモデル化している．

$$P_\theta(w_1, ..., w_N) = \prod_{n=1}^{n} P_\theta(w_n | w_{<n}) = \prod_{n=1}^{N} \prod_{m=1}^{M_n} P_\theta(w_{n,m} | w_{n<m}, w_{<n}) \tag{8}$$

w_n は n 番目の発話を示し，$w_{n,m}$ は n 番目の m 個目の単語を示し，$w_{<n}$ は $w1, ...w_{n-1}$ の省略形である．

VHRED ではさらにこの contextual-RNN に確率的なノイズである潜在確率 z を追加している．この潜在変数 z は以下のように表される．

$$P_\theta(z_n | w_{<n}) = \mathcal{N}(\mu(w_{<n}), \Sigma(w_{<n})) \tag{9}$$

$\mu(w_{<n}$ は現在までの入力文ベクトルの平均，$\Sigma(w_{<n})$ は現在までの入力文ベクトルの分散を表す．そして，$\mathcal{N}(\mu(w_{<n}), \Sigma(w_{<n}))$ はこの平均と分散に従う確率分布を示す．

潜在変数 z は この確率分布に従うサンプリングからえられる値となるため，生成時にランダムなノイズとして働く．具体的には潜在変数 z を用いた次の出力文の予測は以下の式で表される．

$$P_\theta(w_n|z_n, w_{<n}) = \prod_{m=1}^{M_n} P_\theta(w_{n,m}|z_n, w_{<n}, w_{n<m}) \tag{10}$$

VHRED は学習には以下の損失関数から学習される．

$$\log P_\theta(w_1, ...w_N) \geq \sum_{n=1}^{N} -\mathrm{KL}[Q|P] + \mathbb{E}[\log P_\theta(w_n|z_n, w< n)] \tag{11}$$

右辺の $\mathbb{E}[\log P_\theta(w_n|z_n, w< n)]$ は通常のデコーダと同じ生成単語に対する Cross Entropy 誤差を表している．左辺の $\mathrm{KL}[Q|P]$ は現在の入力文の確率分布と次の入力文から求められる確率分布のカルバック・ライブラー距離による誤差を表している．カルバック・ライブラー距離の値が小さいことで 2 つの分布の差を減らし，連続する 2 文の関連性を高くすることを目的としている．

2.2　タスク汎用的に利用される生成モデルの評価手法

本節では雑談対話でも用いられる機械翻訳システムの自動評価手法 BLEU [4] と自動要約システムの自動評価手法 ROUGE [5] について説明する．

2.2.1　BLEU

BLEU [4] は機械翻訳システムの標準的自動評価手法であり，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．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．

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_n \frac{1}{N} \log p_n\right) \tag{12}$$

$$\mathrm{BP} = \begin{cases} 1 & \text{if } \eta > \rho \\ e^{(1-\rho/\eta)} & \text{otherwise} \end{cases} \tag{13}$$

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{\#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \#_g(h_i)} \tag{14}$$

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．

Equation 12 の通り，BLEU での評価の過程で修正 n-gram 精度 p_n を対数に変換するため，p_n が 0 になると計算が不可能になる．このため BLEU には複数の平滑化手法 [18–20] が提案されている．以下ではこれらの平滑化手法について説明する．なおこれらの説明にあたって，m_n をシステム出力と参照出力との共通の n-gram の数，l_n をシステム出力の n-grma の数，これらの平滑化手法による修正後の値を m'_n や l'_n で表す．

まず，平滑化手法 1 を説明する．この手法は Chen らにより提案 [20] されている．この手法では m_n が 0 である場合に，これを ϵ に置き換える手法である．ϵ は経験的に決まるパラメータである．

$$m'_n = \epsilon, \text{if } m_n = 0 \tag{15}$$

次に，平滑化手法 2 を説明する．この手法は Lin らにより提案 [18] されている．この手法では $n > 1$ について各 m_n, l_n に 1 を加算する手法である．

$$m'_n = m_n + 1, \text{for } n \text{ in } 2...N, \tag{16}$$

$$l'_n = l_n + 1, \text{for } n \text{ in } 2...N. \tag{17}$$

次に，平滑化手法 3 を説明する．この手法は NIST (National Institute of Standards and Technology) が提供している BLEU toolkit mteval-v13a[*1]に実装されている手法である．m_n が 0 の時に，$1/invcnt$ に置換され，$invcnt$ は初期値が 1 で，$m_n = 0$ の度に 2 倍になる．

1. $invcnt = 1$
2. for n in 1 to N
3. 　　if $m_n = 0$
4. 　　$invcnt = invcnt \times 2$

[*1] http://www.itl.nist.gov/iad/mig/tests/mt/2009/

5. $m'_n = 1/invcnt$
6. endif
7. endfor

次に，平滑化手法 4 を説明する．この手法は Chen らにより提案 [20] されている．この手法は平滑化手法 3 において m_n が l_n とは無関係に平滑化される点を改善した方法となっている．具体的には平滑化手法 3 の 4 行目について下記の式に置き換える．

$$invcnt = invcnt \times \frac{K}{\ln(len(T))} \tag{18}$$

K は経験的に決まる値，T をシステム出力を表す．これにより，システム出力の文長に沿った平滑化手法となる．

次に，平滑化手法 5 を説明する．この手法も Chen らにより提案 [20] されている．この手法は m_n は m_{n-1} や m_{n+1} に類似するという直感を動機としている．下記の式のように，m_n を自身とその前後の値の平均により平滑化する．

$$m'_0 = m_1 + 1 \tag{19}$$

$$m'_n = \frac{m'_{n-1} + m_n + m_{n+1}}{3} \tag{20}$$

次に，平滑化手法 6 を説明する．この手法は Gao らにより提案 [19] されている．この手法は事前確率 p_n^0 を利用して，p_n の最大尤度推定を補完する手法となっている．事前確率 p_n^0 やそれにより計算される修正後の p_n は以下の通りである．

$$p_n = \frac{m_n + \alpha p_n^0}{l_n + \alpha} \tag{21}$$

$$p_n^0 = p_{n-1} \times \frac{p_{n-1}}{p_{n-2}} \tag{22}$$

α は経験的に決まるパラメタである．この手法では $n > 2$ についてのこの平滑化を行う．

最後に，平滑化手法 7 を説明する．この手法も Chen らにより提案 [20] されている．この手法は平滑化手法 4 と 5 を組み合わせた手法である．まず平滑化手法 4 により計算した後に，さらに平滑化手法 5 を適用する．

以上の平滑化手法について，Chen らによる研究 [20] では平滑化手法 7 が機械学習における人手評価と最も相関が高くなることを示している．

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 [6].

2.2.2　ROUGE

ROUGE [5] 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．つまり BLEU における Equation 14 を下記の式に変更する．

$$p_n = \frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{\#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(r_i)} \#_g(r_i)} \tag{23}$$

ROUGE では派生として，連続した n トークンではなく任意の 2 トークンの組み合わせにより評価する ROUGE-S や係り受け関係を考慮した ROUGE-BE 等が存在する．

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 [6].

第 3 章

関連研究

本章では雑談対話における評価手法について説明を行う．雑談対話タスクでの評価手法は大別して 2 種類存在し，発話と応答の関連性を考慮した機械学習ベースでの自動評価手法と応答多様性を考慮した複数参照に基づく評価手法である．それぞれについて順に紹介する．

3.1 発話-応答関連性を考慮した機械学習ベースの自動評価手法

機械学習ベースの評価手法として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 [21]，並びに人手評価を利用しない教師なし学習により学習した評価関数を利用する評価手法 RUBER [9] を紹介する．

3.1.1 教師データである人手評価に基づく自動評価手法

Adem [21] は Lowe らの提案する，人手評価を教師データとして利用して学習する評価モデルである．Adem では入力発話と参照応答，システム出力を入力し，人手評価を正解データとしてこれを模倣するようなスコアを出力するように評価関数を学習する．学習時に利用する特徴量としてシステム出力と入力発話の類似度、及びシステム出力と参照応答の類似度を利用している．具体的には，参照応答とシステム出力は RNN での最後の隠れ層を参照応答とシステム出力の文ベクトル $r, \hat{r}$ として，入力発話は連続した会話のやりとりも考慮した文ベクトルとして表現するために，階層的 RNN [2] を利用して，その contextual-RNN 側の隠れ層を入力発話の文ベクトル c として変換する。

そしてこの合計値を Equation 24 の計算により，評価値として計算する．

$$score(c, r, \hat{r}) = (c^T M \hat{r} + r^T N \hat{r} - \alpha) / \beta \tag{24}$$

M, N は $\hat{r}$ は行列のパラメータ、 α , β は正規化のための値である．

以上により計算されたモデルによる評価値と人手評価との最小二乗をモデルのロスとして，以下の式により損失関数 L を計算し，パラメータを更新する．

$$L = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human_i]^2 + \gamma \ ||\theta||^2 \tag{25}$$

θ はモデルのパラメータを，K はバッチサイズを，γ はモデルのパラメータに対する正規化定数を表す．

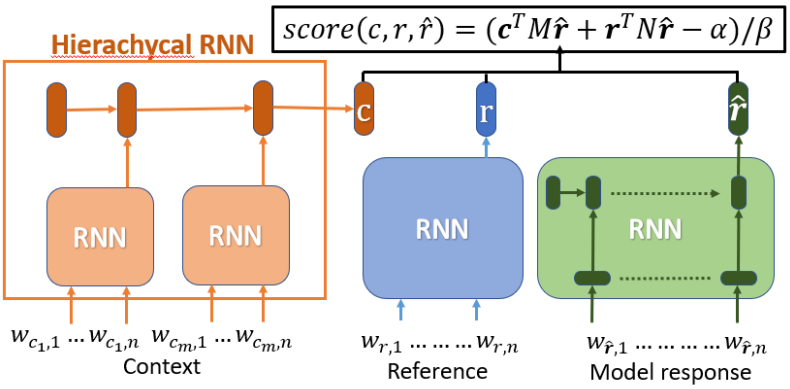

Figure 1　Adem のモデルイメージ

Adem での学習データに必須である人手によるシステム出力に対する評価は非常にコストがかかり、少ないデータセットでの高精度な研究を行うために、階層的 RNN 及び単語の埋め込み表現は事前学習を行っている。具体的には階層的 RNN を利用した対話生成モデルである HRED を人手評価のない単なる会話データで学習することで，階層的 RNN を HRED での Encdoer として事前学習が行える。

Adem の問題点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ことである．

3.1.2　人手評価を必要としない学習ベースの自動評価手法

RUBER [9] 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 Referenced Scorer 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 Unreferenced Scorer 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．本研究では参照応答を利用した 2 種類のモデルによる評価値を組み合わせることで人手評価との高い相関を得た．

具体的に，Referenced Scorer では，参照応答と生成応答の文のベクトル表現を利用するが，文のベクトル表現は事前学習された単語埋め込み表現を以下の式により求める．n 個の単語からなる文の各単語の分散表現を $w_1, w_2, ..., w_n$ として，

$$
\begin{aligned}
x_{max}[i] &= \max(w_1[i], w_2[i], \cdots, w_n[i]) \\
x_{min}[i] &= \min(w_1[i], w_2[i], \cdots, w_n[i]) \\
x &= [m_{max}; x_{min}]
\end{aligned}
$$

以上により求めた参照応答と生成応答の文ベクトル $r, \hat{r}$ の類似度 s_R はコサイン類似度により計算する．

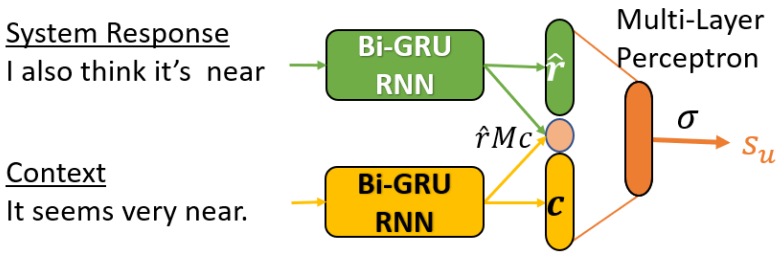

Figure 2　Unreferenced Scorer のイメージ

Unreferenced Scorer では入力発話と生成文の関連性を評価値として計算する．具体的には，まず x をベクトル h に変化する Bidirectional-GRU より入力発話と生成応答

の文ベクトル $c, \hat{r}$ を計算する．

$$
\begin{aligned}
[r_t; z_t] &= \sigma\ (W_{r,z} x_t + U_{r,z} h^{\rightarrow}_{t-1} + b_{r,z}) \\
\tilde{h}_t &= \tanh\left(W_h x_t + U_h (r_t \circ\ h^{\rightarrow}_{t-1}) + b_h\right) \\
h^{\rightarrow}_t &= (1 - z_t) \circ\ h^{\rightarrow}_{t-1} + z_t \circ \tilde{h}_t
\end{aligned}
$$

なお，上記の数式は片方向だけであり，逆向きについての分散表現をまとめて結合したものを文の分散表現とする．

$$
h_t = [h^{\rightarrow}_t, h^{\leftarrow}_t] \tag{26}
$$

次に Equation 27 により結合したベクトルを Multi-layer Perceptron に入力してスカラーの評価値 $s_U(c, \hat{r})$ に変換する．

$$
x = [c; c^T M \hat{r}; \hat{r}] \tag{27}
$$

ただし，M は行列によるパラメータ

Unrefernced Scorer は Equation 28 の損失関数 J によって学習し，パラメータを更新する．

$$
J = max(0, \Delta - s_U(c, r) + s_U(c, r^-)) \tag{28}
$$

ただし Δ はスカラー値のパラメータ、r は参照応答、r^- は負例サンプリングにより選ばれたランダムな参照応答を示す．この学習は、より参照文らしい生成文、つまり妥当な応答に対してはスコア s_U が高くなり、一方で参照文として適当でない、つまり不適切な応答に対してはスコア s_U が低くなることを意図した学習となっている．

最後に以上により求めたそれぞれの評価値 s_R, s_U を組み合わせて最終的なスコアとして利用する．各評価値はその最大値と最小値を利用して，Equation 29 によりそれぞれ $[0, 1]$ に正規化される．

$$
\tilde{s} = \frac{s - \min(s)}{\max(s) - \min(s)} \tag{29}
$$

s は評価値 s_U, s_R のいずれかを意味する．正規化されたスコア $\tilde{s}_R, \tilde{s}_U$ は，これらの最大値・最小値・幾何平均・算術平均のいずれかの組み合わせ方により RUBER 全体としての評価とする．

3.2　複数参照に基づく評価手法

複数参照に基づく評価手法として提案手法の先行研究である複数参照に基づく半自動評価手法 ΔBLEU [7]，並びに人手により応答を拡張し既存の評価手法を複数参照に基づく評価手法へと拡張した Gupta らの研究 [22] を説明する．

3.2.1　ΔBLEU: Discriminative BLEU

ΔBLEU [7] 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 [8] 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 での計算の際に重み付けとして利用することで，より人手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 [8] に倣って BM25 [10] 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（候補発話-候補応答ペア）を収集する．発話-応答ペアの類似度は入力発話と候補発話の類似度，参照応答と候補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複数人による 5 段階のリッカート尺度による評価を $[-1, 1]$ 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，Equation 12 に用いる n-gram 精度 p_n を以下のように計算する．

$$\frac{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_{j: g \in r_{i,j}} \{w_{i,j} \cdot \#_g(h_i, r_{i,j})\}}{\sum_i \sum_{g \in n\text{-grams}(h_i)} \max_j \{w_{i,j} \cdot \#_g(h_i)\}} \tag{30}$$

この式は Equation 14 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．

3.2.2　人手で作成した拡張参照を利用した評価手法

Gupta らは，応答多様性を考慮した評価手法を行うため，DailyDilaog [23] でのテストデータについて複数の参照応答を人手により拡張 [22] を行った．これを利用した評価手法として，表層的類似性に基づく評価手法の BLEU [4] や ROUGE [5] や文の分散表現である Extrema [24] を利用した評価手法などの単一参照応答に基づく評価手法を複数参照応答に基づく評価手法に拡張する提案 [22] を行った．具体的には，システム出力を y，参照応答を r，単一参照に基づく評価手法による評価値を $d(y, r)$ として，複数の参照応答 $R = \{r_1, ..., r_n\}$ を利用して以下のように計算する．

$$score(y, R) = \max_{r_i \in R} d(y, r_i) \tag{31}$$

この複数参照応答を利用した評価手法への拡張により，単一の参照応答による評価より人手評価との相関が向上することを確認した．しかし，この手法では常にテストデータについて複数の参照応答が必要なため，オープンドメインな雑談対話システムを評価するには人手による参照応答のコストが膨大であることが問題点としてあげられる．

第 4 章

提案手法

本節では 3.2.1 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妥当性付き拡張参照応答に自動獲得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．

4.1 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，4.2 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．

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．

4.2　分類期に基づく疑似応答候補の選別と妥当性評価

4.1 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（Figure 3）．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) [25] 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．このような計算

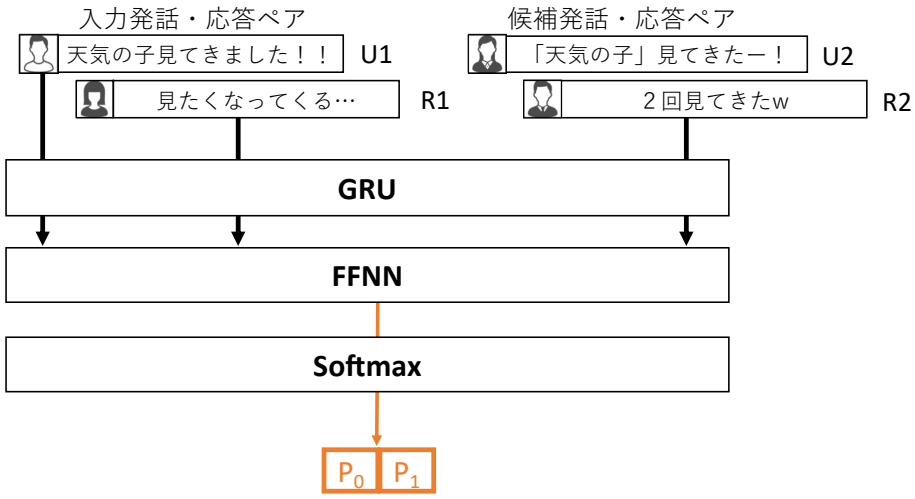

Figure 3　疑似応答の妥当性判定を行う分類器

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．

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）方法や平均値をとる方法，(U1, R1, R2) での分類確率のみを利用する方法の複数が検討可能である．

4.3　妥当性付き複数参照を活用するための評価手法の拡張

提案手法で獲得可能な参照応答には妥当性が付与されている．これを活用するために，Equation 14 に対する Equation 30 のような評価手法への妥当性による重み付けを参考に，Gupta らの提案 [22] 手法である複数参照に基づく評価手法への拡張方法 (Equation 31) に同様の妥当性による重みづけをを考慮した評価手法の拡張方法を提案する．具体的には ΔBLEU の評価手法においての生成応答の参照応答 i との比較の際にその参照応答 i の妥当性で重みを付けることと同様に下記の式による計算により妥

当性付き疑似応答を活用するための評価手法拡張を行う．

$$score(R, r_g) = \max(score(r_i, r_g) \times V_i) \tag{32}$$

r_g は生成応答を，R は拡張参照応答全体を，r_i は拡張参照応答の内一つの応答を，V_i は r_i に付与された妥当性を，$score(r_i, r_g)$ が r_i を利用した単一参照に基づく評価手法の r_g の評価値を示す．

第 5 章

実験

提案手法である妥当性付き複数参照応答の有効性を確認する．このためにまず，ΔBLEU での人手評価による妥当性付与した拡張参照応答との提案手法の比較を予備実験として行い，BLEU/ΔBLEU に関して提案手法の有効性を確認する．次に，提案手法による妥当性付き複数参照応答の有効性を，妥当性を活用できる評価手法の拡張手法を利用して確認する．

5.1 予備実験：妥当性付き複数参照応答の獲得における先行研究との比較

本章では第 4 章で説明した ΔBLEU における疑似応答の収集方法，並びに妥当性評価の自動化についての提案手法を，それぞれ既存手法との比較実験を行った．

5.1.1 実験設定

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．

5.1.1.1 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[*2] 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

[*2] https://dev.twitter.com/overview/api

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[*3]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タセットから削除した．他に以下の前処理を順に行った．

- Twitter API の仕様により，115 文字以上の投稿は省略されるため，省略された投稿の削除
- “@” から始まるユーザー名，“http” から始まる URL のトークン化
- 絵文字の削除[*4]
- 応答において投稿の先頭に出現するユーザートークンの削除
- NFKC 正規化
- 大文字を小文字に変換
- 一文字の 4 回以上の繰り返しを 3 回に省略
- 二文字の 3 回以上の繰り返しを 2 回に省略
- アルファベットまたは数字または空白の占める割合が，投稿内容の 6 割以下の

[*3] https://github.com/shuyo/ldig
[*4] https://pypi.org/project/emoji/

投稿の削除

- 5 文字以下の投稿の削除
- 同日内で投稿内容が重複する投稿はすべて bot とみなし，削除

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 万対を収集し，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．最終的に，構築された対話データは基本的に SentencePiece [26] を用いてトークン化を行った．SentencePiece によるトークン化のために，文字分割のモデルを学習する必要があるが，基本的に各訓練データにより，語彙数 16,000 はそれ以外はデフォルトのパラメータで学習した．ただし，上記データ約 5,000 万対をデータベースとして利用した際には，1,000 万対を学習データとしてランダムにサンプルを行い学習した．以下より，雑談対話生成モデル・応答妥当性付与のための分類器について順に説明する．

5.1.1.2　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[*5]で実装されたテキスト生成ライブラリ fairseq[*6]の Transformer [12] 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 Table 1 に示されるもの以外は元文献 [12] と同一の値に設定した．

雑談応答モデル（および自動評価）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 [7] に倣って 5 段階のリッカート尺度による評価を行った．

[*5] https://pytorch.org/
[*6] https://ai.facebook.com/tools/fairseq/

Table 1　ハイパーパラメタ設定

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
	学習率	0.001
	最適化方式	Adam [27]
		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

5.1.1.3　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．Table 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．また分類器により自動付与する妥当性には 2 種の 2 値分類確率のうち最も大きい値を利用した．

5.1.1.4　比較手法

本実験では，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．

なお実験での提案手法で利用する文の分散表現は，予備実験の結果をもとに，WordPiece [28] によってトークン化を行い，事前学習された BERT [29] 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手法により得ることとした．

まず，疑似応答の収集手法に関して類似度の計算対象として，発話のみに基づく類似度の計算と発話と応答それぞれの類似度の積による発話-応答としての類似度の計算を比較する．同時に類似度計算手法として BM25 による表層的類似度による計算と，BERT による単語埋め込みを平均化して文ベクトルとみなし，このコサイン類似度により類似度の計算を行う分散表現に基づく計算を比較した．このとき，発話-応答ペアの類似性を BM25 で計算する手法が既存収集手法 [8]，発話のみの類似性を分散表現ベースの手法で計算する手法が提案手法となっている．以上のそれぞれの方法により疑似応答を収集し，得られた疑似応答を BLEU での評価に直接用いて，ど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応答として用いて BLEU を計算する [8] 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 [7] 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 に倣って Kendall の τ および Spearman の ρ を利用した．

5.1.1.5　評価手順

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 [7] 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[*7]によりトークン化を行った．

5.1.2　実験結果

Table 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 が人手評価との相関に関して帰無仮説を棄却できない，つまり有意な相関が存在しないことと，疑似応答の収集方法として提案手法が既存手法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，Table 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．

Table 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．

[*7] https://github.com/stanfordnlp/stanfordnlp

5.1.3　予備実験のまとめ

本実験では大規模な 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既存手法と比較することで提案手法の有効性を確認した．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．

Table 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

参照応答	類似対象	類似度計算	Kendall's τ	(p-value)	Speaman's ρ	(p-value)
single			-0.012	(0.88)	-0.019	(0.85)
all	発話と応答	BM25	0.109	(0.14)	0.147	(0.14)
all	発話のみ	BM25	0.123	(0.10)	0.169	(0.09)
all	発話と応答	分散表現	0.120	(0.11)	0.158	(0.12)
all	発話のみ	分散表現	**0.227**	(< 0.01)	**0.297**	(< 0.01)

Table 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）

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)

5.2　本実験：提案手法を活用した拡張評価手法の従来手法との比較

本節では第 4 章で説明した提案手法により獲得可能な妥当性付き複数応答を利用して，単一の参照応答に基づく自動評価手法を妥当性付き複数参照応答に基づく評価手法へ拡張することの有効性を確認した．

5.2.1　実験設定

本節では，Twitter 上の大規模日本語対話データセットを用いて実験を行う．

5.2.1.1　大規模日本語データセット

実験で利用する大規模日本語データセットは基本的に予備実験と同様の手順により収集を行った．但し，前処理の過程で

- “http” からなる URL のトークン化
- アルファベットまたは数字または空白の占める割合が，投稿内容の 6 割以下の投稿の削除

の箇所を

- “http” から始まる URL と “#” から始まるハッシュタグを含む投稿の削除
- 漢字またはひらがなまたはカタカナの占める割合が，投稿内容の 4 割以下の投稿の削除

に変更，また

- 自身の投稿へメンションを行った投稿の削除

を追加した．この URL もしくはハッシュタグを含む投稿の削除の理由は，Twitter に特徴的な情報や，言語による情報以外による要素を排除することで，一般的なテキスト上での雑談対話データとしての構築を目的とするためである．雑談対話応答システムの学習データ並びにテストデータの構築の際には，5.1.1 節と同様に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行った投稿を応答とする

が，応答に発話と同一のアカウントからのメンションが行われた発話-応答ペアのみを抽出し，雑談対話データセットを構築した．これは Twiiter データから会話データを構築した研究 [14,30] を参考に，雑談対話は 2 者間の間で行われ，かつ 1 発言ずつだけのやりとりで終わらない会話データの構築を行うためである．その他の雑談対話データとしての利用では，5.1.1 節と同様との方法で雑談対話データセットを構築した．分類器の学習データの学習のための雑談対話データセットではできる限りデータ数を増やすため，応答が 4 つ以上存在する発話をすべて削除するのではなく，応答が複数存在する発話については高々 3 つまでの応答とのペアを利用した．その他の雑談対話データとしての利用の際は，発話に対して複数の応答が存在した場合，基本的にランダムに 1 つを選んだ．

5.2.1.2　雑談対話応答システム

提案手法評価のために，評価対象となる雑談対話応答システムの学習を行った．実験では 2.1 節で説明した，2.1.1 節の BM25 [10] を利用した検索型モデルと 2.1.2 節の VHRED [11] による生成型モデルをそれぞれ利用した．BM25 の実装にはオープンソースライブラリ gensim[*8]により，VHRED の実装には著者らにより GitHub 上で公開されているコード[*9]を利用した．VHRED の学習データは 2018 年内に投稿された対話データから 240 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．BM25 のデータベースは VHRED の学習データをそのまま利用した．これらの対話データは SentencePiece を用いてトークン化を行った．SentencePiece によるトークン化のために，文字分割のモデルを学習する必要があ るが，訓練データにより，語彙数 32,000，それ以外はデフォルトのパラメータで学習しトークン化を行った．VHRED 並びに BM25 のハイパーパラメタは基本的にその実装の初期値のままの値に設定した．ただし，VHRED のバッチサイズは 40 とした．

雑談対話応答モデル（および自動評価）の評価データの構築のために，2019 年 1 月から 9 月までの雑談対話データのうち応答が 2 つ以上存在する発話 47,361 件を抽出した．生成型雑談対話モデルは特定のトークンの繰り返しが多い応答[*10]を生成すること

[*8] https://radimrehurek.com/gensim/
[*9] https://github.com/cgsdfc/HRED-VHRED
[*10] 実例として「野菜は野菜を選ぶと野菜が野菜を選ぶので野菜は野菜です.」などを生成した

があり，これは応答として破綻していることが多い．このような生成応答は評価データおける評価値を低い方に偏らせる原因となるため，このような応答を生成した会話を除外した．具体的には，VHRED の各生成応答ごとに文中に出現する頻度上位 3 トークンで 2 回以上出現したトークンが「トークンの長さ＋頻度」の尺度で 5 以上（つまり 1 文字のトークンが 4 回以上出現，3 文字以上のトークンが 2 回以上出現など）となる応答を生成した会話を除外した．以上のクリーニング済み評価データ 26,995 発話から 400 の発話をランダムに選んだ．この発話に対し，訓練データからそれぞれ BM25 により発話の類似度が最も高くなった発話の応答を検索型モデルによる応答として抽出した．ただし，BM25 の性質上，類似度判定対象の文中（ここでは評価データの発話文）に出現するトークンが高頻度に出現するデータベース中の文を最類似文として採択しやすく，一方で Twitter 上でのこのような文が出現すること[*11]がありこの投稿は会話での発話文として適切でないため取り除く必要がある．具体的には，発話内の 2 回以上出現するトークンの合計単語数がその発話の全文字数の 5 割以上を占める発話，もしくは同じトークンが 5 回以上出現する発話を含む会話 (計 35131 ペア) をデータベースとして利用する訓練データから除外した．VHRED による生成応答と，BM25 によりデータベースから抽出した応答，そして実応答の一つを理想的なモデルによる生成文とみなして，各発話の生成応答を 3 種類用意した．なお，各発話の他方の実応答を正解参照として利用する．

この 400 発話に対する 3 種類の応答の計 1,200 ペアについて，著者らの所属する研究室の学生 6 人により 5 段階のリッカート尺度により発話に対する生成応答の適切さの評価を行った．具体的には，「発話・応答が会話として適切に成立するかどうかのみ評価し、道徳的な観点で適切な応答かは考慮しない」，「会話が成立するかの判定には、常識的な範囲で適当な背景知識を仮定しても良い」といった注意書きの下で，評価値 5 で十分に適切な応答，もしくは実際の応答だと思えるもの，評価値 4 でどちらかというと適切な応答，評価値 3 で適切かどうか判断に迷う応答，評価値 2 でどちらかというと不適切な応答，評価値 1 で応答としてふさわしくない，もしくは発話に対する応答として理解できない応答というガイドラインにより評価を行った．実験で利用する際には 1,200 ペアを 600 ペアずつの開発データとテストデータに分割した．

[*11] 例えば生クリーム食べたい生クリーム食べたい生クリーム食べたい」のように文や「スクールアイドルアクティビティビティビティビティビティビティ」のように一部の文字列を繰り返した投稿が存在した

5.2.1.3　妥当性応答分類器の学習

次に応答の妥当性評価のための分類器の学習を行った．2017 年の対話データから分類器の学習データとして 280 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．学習に利用したハイパーパラメタは予備実験での Table 1 と同じ値を設定した・同様に，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．なお，分類器の分類確率として利用する手法はハイパーパラメタとして開発データにより選択をした．

5.2.1.4　比較手法

本実験では単一参照応答に基づく複数の評価手法を通じて．提案手法による妥当性付き複数参照応答の有効性を確認する．

まず疑似応答の収集手法に関して予備実験と同様の手順により，日本語でも同様の結果となるか確認を行う．同時に，分散表現を利用した疑似応答の収集手法について複数の手法を検討する．具体的には，トークン化手法並びにそのベクトル化手法として

- MeCab[*12]・Glove [31]
- SentencePiece [26]・Glove [31]
- Mecab+WordPiece [28]・BERT [29]

について人手評価との相関に関して比較した．なお，BERT の利用の際には Wolf らにより実装 [32] された BERT モデルに移植された，日本語の Wikipedia をコーパスとして学習した BERT モデル[*13]を利用した．また Glove の元文献 [31] での考察において，意味類推タスクでは単語の埋め込み表現を学習する際の前後の単語数（ウィンドウサイズ）が影響することが確認されている．このため Glove による単語の分散表現を利用する際には，複数のウィンドウサイズにより学習した Glove を利用した収集方法についても比較した．

[*12] https://taku910.github.io/mecab/
[*13] https://github.com/cl-tohoku/bert-japanese

次に表層的類似性に基づく評価手法として，予備実験と同様に，単一参照による BLEU と複数参照による BLEU，分類確率を妥当性評価として利用した ΔBLEU-auto を比較する．2.2.1 節で説明したように，BLEU [4] には複数の平滑化手法 [18–20] が存在する．これらの平滑化手法はハイパーパラメタとして開発データにより選択した．

さらに，分散表現ベースでの評価手法についても提案手法による妥当性付き複数参照応答の有効性を確認する．具体的には，文における単語の埋め込み表現の平均化 (Average)・極値化 (Extrema) [24]・それぞれの要素の最大をとったベクトルと最小をとったベクトルの結合ベクトル (Maxmin) [9] でそれぞれ提案手法の有効性を確認した．なお，この評価の手法での単語の分散表現は，MeCab によりトークン化し，ウィンドウサイズ 5，ベクトルサイズ 50 で学習した Glove を利用した．

5.2.2　RUBER との組み合わせ

3.1.2 節で説明したように，自動評価手法 RUBER は入力発話と生成応答の関連性を学習モデルにより推定する Unreferenced Scorer と，参照応答と生成応答を分散表現でのコサイン類似度により計算する Referenced Scorer を組み合わせることで生成応答を評価している．Referenced Scorer は他の参照応答に基づく評価手法と置換が可能であるため，提案手法を利用した評価手法についても同様である．実験においては，提案手法の有効性を確認した後に，RUBER の Unreferenced Score との組み合わせた評価手法について検討する．RUBER の Unreferenced Score は学習する必要があり，分類器の学習に利用した学習・開発データにより学習を行った．学習時のハイパーパラメタを Table 4 に示す．Unreferenced Scorer 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．

5.2.3　評価手順

評価データに対する疑似応答候補は，2017 年の対話データセット中で発話投稿の重複しない発話-応答ペア約 1600 万対から収集を行った．まず，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 2000 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，入力応答に対する応答としての妥当性を付与した．ただし，ΔBLEU とは異なり，

Table 4　Unreferenced Scorer (RUBER) のハイパーパラメタ設定

学習率	10^{-5}
最適化方式	Adam
	$\beta_1 = 0.9$
	$\beta_2 = 0.999$
Δ	0.075
最大エポック数	20
バッチサイズ	1000

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．評価付き入力発話-生成応答ペアの開発用データ 600 対を利用して，最適な疑似応答件数，疑似応答選別の際の閾値，分類器の利用する確率出力方法，BLEU における平滑化手法を調整した．ただし，分散表現を利用した評価手法では計算コストが膨大になるため類似度上位 100 件までの中で最適な値を探索した．開発データで最も人手評価 Spearman の順位相関係数 ρ の最も高い設定をテスト時で利用した．RUBER の Unreferenced Scorer と組み合わせた評価手法について表層的類似性に基づく評価手法と，分散表現を利用した評価手法のそれぞれで最も性能の高い手法についてのみ検討した．この際，3.1.2 節で述べたように，組み合わせ方には複数の手法が存在するため，これについても開発データにより最適な手法を調整し，テスト時のその手法により人手評価との相関を評価した．なお，評価手法の評価の際には，MeCab によりトークン化を行った．BLEU による評価では BLEU-2 での評価を行う．このため平滑化手法 6 は平滑化を行わないことと同義のため利用しない．

5.3　実験結果

5.3.1　雑談対話生成応答に対する人手評価の結果

評価値のアノテータごとの分布を Table 5 に示す．評価値 3 を基準に対称的な分布になっており，全体として適切な応答のみもしくは不適切な応答のみに偏っていないことが確認できる．

評価者間の評価の合意関係を Table 6 に示す．雑談対話評価手法に関する研究 [6,21]

Table 5　各評価者の評価値分布

アノテータ	評価値				
	1	2	3	4	5
1	535	100	86	146	333
2	407	106	77	85	525
3	478	59	57	89	517
4	677	65	35	110	313
5	320	217	78	262	323
6	320	156	117	197	410
平均	456.2	117.2	75	148.2	403.5

Table 6　評価者間の合意関係

手法	κ
Cohen　max	0.419
Cohen　min	0.304
Fleissen	0.369

での一致度の尺度として一般的に使われる，個人間での一致度を測る Cohen の κ [33]，並びに一般的に複数者間での一致度の測定に利用される Fleissen の κ [34] によりアノテーションデータを評価した．Ladis と Koch ら [35] によると $0 \leq \kappa \leq 0.2$ で “slight”，$0.2 \leq \kappa \leq 0.4$ で “fair” であり，雑談対話応答の評価手法に関する研究 [6, 21] では Cohen の $\kappa = 0.2$ を閾値とすることが一般的なため，実験ではすべての評価値を利用した．

Cohen もしくは Fleissen の κ による評価者間の評価の一致度を計測する方法は，主にカテゴリ分類タスクを対象にしている．一方で，この発話に対する生成応答の適切さの評価はリッカート尺度による順序尺度である．そこで各アノテータ間の評価値の順位相関を Spearmanρ，Kendallτ により計測した結果を Table 7 に示す．

Table 7　個人間の評価の順位相関

	Spearmanρ	Kendallτ
max	0.756	0.666
average	0.697	0.613
min	0.651	0.569

多岐にわたる分野で利用される Guilford らの相関係数の指標 [36] では相関係数が 0.2 以下で “slight”，0.2 以上 0.4 以下で “low”，0.4 以上 0.7 以下で “moderate”，0.7 以上 0.9 以下で “high” としている．このため，個人間の相関に関しては “moderate” または “high” であると言える．

リッカート尺度は順序尺度であるため，複数のアノテータによる評価を集約する場合，中央値や最頻値を利用することが望ましい．しかし，5 段階でのリッカート尺度による評価についてその中央値または最頻値を利用した場合，高々 9 段階の評価値しか存在せず，多くの評価対象がタイとなる．これを避けるため，評価値を間隔尺度とみなし，複数のアノテータによる評価をその平均により集約した．実際の評価値について 2 群に分け，それぞれの平均値によって集約した際の 2 群の相関を Pearson の積率相関係数 r，Spearman の順位相関係数 ρ，Kendall の順位相関係数 τ により計算した分割した 2 群の人数ごとにまとめた結果を Table 8 に示す．

Table 7 と Table 8 の 1 人対 4 人の相関係数を比較して，複数人の評価値を平均により集約することで，特定の個人による評価との相関が向上することが確認できる．また Table 8 での 2 群に分けた際の人数比ごとの比較をすると，人数差が少ないほど群間での相関係数が向上することが悪人できる．これは両群の少ない方の人数について，群内での評価値の平均化により個人の評価対象ごとの偏りが減少したことが原因ではないかと考えられる．自動評価手法と比較する際には人手による評価値は全員の評価値を平均化した値を利用する．以上の結果より，アノテータ 6 人全員の評価値を平均化することで，理想的な評価手法では Table 8 における 3 人対 3 人以上の相関係数を期待できる．

Table 8　評価値についての群間での相関

相関係数		1 人対 5 人	2 人対 4 人	3 人対 3 人	全体
Pearson r	max	0.840	0.879	0.882	0.882
	average	0.794	0.855	0.868	0.853
	min	0.754	0.828	0.853	0.754
Spearman ρ	max	0.840	0.872	0.876	0.876
	average	0.796	0.857	0.869	0.854
	min	0.764	0.834	0.860	0.764
Kendall τ	max	0.712	0.737	0.742	0.742
	average	0.667	0.720	0.732	0.718
	min	0.631	0.693	0.721	0.631

5.3.2　複数参照に基づく BLEU よる疑似応答収集手法の比較

まず，提案手法による分散表現を利用した疑似応答収集に関して，利用する分散表現に関して比較を行った．具体的には，トークン化，ベクトル化の手法として

- トークン化：MeCab，ベクトル化：Glove
- トークン化：SentencePiece，ベクトル化：Glove
- トークン化：MeCab + WordPiece，ベクトル化：BERT

を比較した．それぞれの手法を利用して，BLEU の複数参照応答として利用した際に，開発データにおいて人手評価との相関により Spearman の相関係数 ρ により調整した結果，並びにそのパラメータでテストデータについて評価した結果を Table 9 に示す．開発データにおける比較として，Pearson の相関係数 r では Sentencepiece によるトークン化と Glove によるベクトル化が人手評価ろの最も高い相関を得ているが，それ以外では Mecab によるトークン化と Glove によるベクトル化が最も高い相関となった．実験では，開発データにおいて Spearman の相関係数 ρ のスコアにより調整を行うため，以後の実験では Mecab によるトークン化と Glove によるベクトル化を採用した．

Table 9　分散表現を利用した発話に基づく疑似応答収集時の分散表現での比較（p 値は全て 10^{-3} 以下）

データ	トークン化 ベクトル化	Mecab Glove	SentencePiece Glove	MeCab・WordPiece BERT
開発	Pearson r	**0.390**	0.392	0.329
	Spearman ρ	**0.386**	0.379	0.326
	Kendall τ	**0.272**	0.268	0.231
テスト	Pearson r	0.274	0.243	0.181
	Spearman ρ	0.263	0.232	0.176
	Kendall τ	0.184	0.163	0.122

Pennington らの Glove による単語の分散表現を利用する場合，元文献 [31] で述べられているようにタスクごとに最適な学習時のウィンドウサイズが存在する．これを考慮し，複数のウィンドウサイズで学習した Glove でそれぞれ学習データの値から最適なパラメータの調整を行った．Mecab によりトークン化し，それぞれのウィンドウサイズによる Glove の単語の分散表現を利用して疑似応答を収集し，BLEU の複数参照応答として利用した際に，開発データにおいて人手評価との相関により Spearman の相関係数 ρ により調整した結果，並びにそのパラメータでテストデータについて評価した結果を Table 10 に示す．開発データにおける比較として，ウィンドウサイズ 10 で学習した Glove が最も高い相関となった．以後の実験ではウィンドウサイズ 10 で学習した Glove によるベクトル化を採用した．

最後に，予備実験と同様に，先行研究である ΔBLEU [7] での疑似応答の収集手法である BM25 を利用した収集手法と以上までに述べた分散表現を利用した収集手法の比較，同時に発話-応答の両方において類似度を計算する収集手法と提案手法である発話のみの類似度に着目した収集手法を比較する．それぞれの手法で疑似応答を収集し，BLEU の参照応答として利用した際に，開発データにおいて人手評価との相関により Spearman の相関係数 ρ により調整した結果，並びにそのパラメータでテストデータについて評価した結果を Table 11 に示す．開発データにおける比較により，5.1 節で

Table 10　ウィンドウサイズを変えた際の比較（p 値は全て 10^{-4} 以下）

	ウィンドウサイズ	5	10	20	50
開発	Pearson r	0.390	**0.398**	0.375	0.382
	Spearman ρ	0.386	**0.395**	0.373	0.381
	Keandall τ	0.272	**0.278**	0.262	0.267
テスト	Pearson r	0.274	0.266	0.269	0.300
	Spearman ρ	0.263	0.258	0.261	0.279
	Keandall τ	0.184	0.180	0.182	0.197

の実験と同様に，提案手法である分散表現を利用した，発話のみの類似性に基づく疑似応答の収集が人手評価との相関における比較により疑似応答の収集手法としてふさわしいことが確認できた．以後の実験では同様の手法を利用した．

Table 11　疑似応答の収集方法の変化と BLEU による評価と人手評価との相関（p 値は全て 10^{-3} 以下）

	類似度計算対象	発話と応答	発話のみ	発話と応答	発話のみ
	類似度計算手法	BM25	BM25	分散表現	分散表現
開発	Pearson r	0.290	0.327	0.323	**0.398**
	Spearman ρ	0.294	0.320	0.338	**0.395**
	Keandall τ	0.204	0.226	0.238	**0.278**
テスト	Pearson r	0.164	0.177	0.192	0.266
	Spearman ρ	0.138	0.179	0.191	0.258
	Keandall τ	0.096	0.124	0.133	0.180

5.3.3　複数参照応答に基づく BLEU と ΔBLEU の比較

5.1 節での実験と同様に，単一の参照応答のみを利用した BLEU による評価 (BLEU-single) と拡張参照応答を利用した BLEU による評価 (BLEU-multi) と拡張参照応答と分類器による分類確率を妥当性として利用した提案手法による評価 (ΔBLEU-auto) を比較する．以上までの結果から，MeCab によりトークン化し，ウィンドウサイズ 10 で学習した Glove を利用して疑似応答を収集し，学習済み分類器を利用して，妥当性を付与した．この疑似応答を利用した BLEU-muti とさらに自動付与した妥当性を利用する ΔBLEU-auto，そして単一の参照応答による BLEU-single をそれぞれ開発データで調整を行い，その開発データで人手評価との Spearman の相関関係数 ρ について最も高かった相関係数と，そのパラメータにより評価したテストデータでの相関係数の結果をそれぞれ Table 12 に示す．開発データ・テストデータの両方で複数参照を利用した評価手法 (BLEU-multi, ΔBLEU-auto) での評価が単一参照での評価 (BLEU-single) より人手評価との相関に関して向上していることが確認できる．BLEU-multi と ΔBLEU-auto を比較すると開発データでは BLEU-multi による評価ががより高い相関となっているが，テストデータではこれが逆転している．これは分類確率による妥当性を利用することで，異なるデータに対しての評価にも柔軟に対応できるデータに対する頑健性が備えられたのではなかと推測される．

Table 12　提案手法を利用した BLEU による評価の段階的な比較（p 値は全て 10^{-3} 以下）

	評価手法	BLEU-single	BLEU-multi	ΔBLEU-auto
開発	Pearson r	0.296	**0.398**	0.394
	Spearman ρ	0.207	**0.395**	0.390
	Keandall τ	0.144	**0.278**	0.274
テスト	Pearson r	0.264	0.266	**0.278**
	Spearman ρ	0.255	0.258	**0.270**
	Keandall τ	0.176	0.180	**0.189**

5.3.4　複数参照応答に基づく分散表現を利用した評価

以上までの方法により収集した疑似応答を利用して，複数参照応答を利用した分散表現による評価についても比較する．具体的にはウィンドウサイズ 5 で学習されたベクトルサイズ 50 の Glove による単語の分散表現を平均化して評価に利用する方法 (Average) と極値化して評価に利用する方法 (Extrema) [24] と，最大・最小ベクトルを結合して評価に利用する方法 (Maxmin) [9] のそれぞれを単一参照のみを利用した場合と，複数参照を利用した場合と，さらに妥当性を利用した場合で比較する．単一参照を利用した場合の評価手法は最後に “-single” を，複数参照を利用した場合の評価手法では最後に “-multi” を，複数参照とそれに自動付与された妥当性を利用した評価手法では最後に “-delta” を付けて示す．以上までの実験と同様に開発データで調整した結果とテストデータでの結果をそれぞれ Tables 13〜15 に示す．分散表現を平均化した手法では，テストデータにおいてすべて p 値が 0.01 を上回るため，無相関である仮説，帰無仮説を棄却できなかったが，いずれの場合についても，テスト時について，単一参照応答を利用した評価手法よりも，複数参照応答を利用した場合の人手との相関が向上した．さらに妥当性を利用した評価が最も人手評価との高い相関となることを確認できる．以上により，複数の参照応答に基づく評価手法においても，提案手法による疑似応答の収集と，疑似応答へ自動付与したの妥当性を評価に利用する有効性を確認できた．

Table 13　分散表現 Average による評価．() 内に p 値を併記する．

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	Average-sngle	0.054 (0.190)	0.172 (<0.001)	0.121 (<0.001)
	Average-multi	0.093 (0.023)	0.179 (<0.001)	0.124 (<0.001)
	Average-delta	0.062 (0.128)	0.183 (<0.001)	0.127 (<0.001)
テスト	Average-single	0.026 (0.529)	0.043 (0.288)	0.029 (0.304)
	Average-multi	0.032 (0.432)	0.058 (0.156)	0.038 (0.156)
	Average-delta	0.045 (0.272)	0.068 (0.098)	0.044 (0.116)

Table 14　分散表現 Extrema による評価（p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	Extrema-sngle	0.237	0.257	0.177
	Extrema-multi	**0.369**	**0.379**	**0.267**
	Extrema-delta	0.358	0.368	0.260
テスト	Extrema-single	0.171	0.217	0.148
	Extrema-multi	0.245	0.277	0.193
	Extrema-delta	**0.260**	**0.288**	**0.200**

Table 15　分散表現 Maxmin による評価．() 内に p 値を併記する．

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	Maxmin-sngle	0.076 (0.063)	0.196 (<0.001)	0.136 (<0.001)
	Maxmin-multi	0.172 (<0.001)	0.269 (<0.001)	0.185 (<0.001)
	Maxmin-delta	**0.178 (<0.001)**	**0.271 (<0.001)**	**0.186 (<0.001)**
テスト	Maxmin-single	0.070 (0.087)	0.096 (0.019)	0.065 (0.020)
	Maxmin-multi	0.123 (0.003)	0.162 (<0.001)	0.110 (<0.001)
	Maxmin-delta	**0.149 (<0.001)**	**0.182 (<0.001)**	**0.123 (<0.001)**

5.3.5　Unreferenced Scorer と組み合わせた評価手法の比較

提案手法により収集した疑似応答とその分類器により自動付与した妥当性を利用した際の，複数参照応答による表層ベースでの評価手法 ΔBLEU-auto と分散表現ベースでの評価手法 Extrema-delta を RUBER の Refernced Scorer (RS)，つまり Maxmin による分散表現ベースでの単一の参照応答による評価と置換した際の評価について，比較を行った．開発データでは，RUBER での Blender つまり Referenced Scorer (RS) と Unreferenced Scorer (US) による評価値の組み合わせ方についての調整も行った．開発データでの Spearman の相関係数 ρ に基づく最適な人手評価と

の相関の結果と，そのパラメータでのテストデータの結果を Table 16 に示す．開発データでは Unreferenced Scorer と ΔBLEU を組み合わせた評価手法が最も高い相関となったが，テストデータでは Pearon の積率相関係数 r では同様に Unreferenced Scorer と ΔBLEU の組み合わせが，Spearman'sρ と Kendall τ の順位相関係数については Unreferenced Scorer と Extrema-delta が最も高い相関となったことが確認できる．これは Unreferenced Scorer と組み合わせない場合でも同様の結果となっている．RUBER のオリジナルの手法では単一参照に基づく Maxmin による分散表現を利用した評価手法が，テストデータについて帰無仮説を棄却できなく実質無相関であるため，Unreferenced Scorer より低い結果となることが確認できる．結果，RUBER での Unreferenced Scorer と組み合わせた場合についても，同様に妥当性付き複数参照を利用した評価手法が有効であることを確認できた．

Table 16　RUBER と組み合わせた場合の比較（() 内に p 値を併記する．ただし表記されない場合，p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	US	0.393	0.401	0.280
	RS	0.076 (0.063)	0.196	0.136
	ΔBLEU-auto	0.394	0.390	0.274
	Extrema-delta	0.358	0.368	0.260
	US+RS (RUBER)	0.373	0.395	0.276
	US+ΔBLEU-auto	**0.496**	**0.511**	**0.364**
	US+Extrema-delta	0.471	0.484	0.342
テスト	US	0.334	0.364	0.255
	RS	0.070 (0.087)	0.096 (0.019)	0.065 (0.020)
	ΔBLEU-auto	0.278	0.270	0.189
	Extrema-delta	0.260	0.288	0.200
	US+RS (RUBER)	0.318	0.356	0.249
	US+ΔBLEU-auto	**0.398**	0.420	0.296
	US+Extrema-delta	0.390	**0.431**	**0.304**

5.3.6　考察

Table 12 での評価は開発データでの調整の際に，BLEU における平滑化手法についても開発データで最も相関の高い手法を採択している．この考察では，同様に得られた疑似応答と自動付与した妥当性を利用して，それぞれの平滑化手法ごとに，以上までの実験と同様に開発データでの Spearman'ρ に基づく最適な人手評価との相関の結果と，そのパラメータでのテストデータの結果を Tables 17〜22 に示す．単一参照応答に基づく評価手法 BLEU-snigle による評価では人手評価との相関について，開発データと同様にテストデータでも平滑化手法 2 または 4 が比較的高い結果となっており，この手法が適切であるように推測される．妥当性を利用しない複数参照応答に基づく評価手法 BLEU-multi については，Pearson r では BLEU-single と同様に，開発データとテストデータで人手評価との相関が最も高い状態のままであるが，Spaerman's ρ や Kendall τ の順位相関については開発データでは相関の低い平滑化手法 5 や 7 がテストデータで同等の結果となっている．そして，妥当性を利用する複数参照応答に基づく評価手法 ΔBLEU についてはいずれの相関係数についても開発データで最適な平滑化手法である 2 や 4 とは異なる 5 や 7 がテストデータで高い相関となっている．以上の考察から，各手法 BLEU-single, BLEU-multi, ΔBLEU でそれぞれ最適な平滑化手法が異なるのではないかと推測される．

Table 17　平滑化手法 1 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.277	0.191	0.133
	BLEU-multi	0.338	0.328	0.228
	ΔBLEU-auto	0.334	0.322	0.223
テスト	BLEU-single	0.222	0.189	0.131
	BLEU-multi	0.212	0.194	0.134
	ΔBLEU-auto	0.223	0.206	0.142

Table 18　平滑化手法 2 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.289	0.193	0.134
	BLEU-multi	0.398	0.395	0.278
	ΔBLEU-auto	0.394	0.390	0.274
テスト	BLEU-single	0.268	0.262	0.182
	BLEU-multi	0.266	0.258	0.180
	ΔBLEU-auto	0.278	0.270	0.189

Table 19　平滑化手法 3 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（() 内に p 値を併記する．ただし表記されない場合，p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.268	0.179	0.136
	BLEU-multi	0.326	0.319	0.222
	ΔBLEU-auto	0.369	0.367	0.256
テスト	BLEU-single	0.208	0.133 (0.002)	0.100 (0.002)
	BLEU-multi	0.203	0.184	0.129
	ΔBLEU-auto	0.252	0.233	0.162

Table 20　平滑化手法 4 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.296	0.207	0.144
	BLEU-multi	0.388	0.388	0.275
	ΔBLEU-auto	0.386	0.385	0.271
テスト	BLEU-single	0.264	0.255	0.176
	BLEU-multi	0.260	0.251	0.176
	ΔBLEU-auto	0.273	0.264	0.184

Table 21　平滑化手法 5 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（() 内に p 値を併記する．ただし表記されない場合，p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.255	0.174	0.140
	BLEU-multi	0.250	0.289	0.209
	ΔBLEU-auto	0.353	0.373	0.265
テスト	BLEU-single	0.167	0.085 (0.038)	0.067 (0.041)
	BLEU-multi	0.202	0.251	0.180
	ΔBLEU-auto	0.301	0.345	0.242

Table 22　平滑化手法 7 を利用した際の BLEU-single，BLEU-multi，ΔBLEU の比較（() 内に p 値を併記する．ただし表記されない場合，p 値は全て 10^{-3} 以下）

データ	評価手法	Pearson r	Spearman ρ	Keandall τ
開発	BLEU-single	0.255	0.174	0.140
	BLEU-multi	0.250	0.289	0.209
	ΔBLEU-auto	0.346	0.369	0.261
テスト	BLEU-single	0.167	0.085 (0.038)	0.067 (0.041)
	BLEU-multi	0.202	0.251	0.180
	ΔBLEU-auto	0.279	0.345	0.241

Figure 4 に，利用する最大疑似応答数に対する各閾値で選別した際の疑似応答の利用割合を示す．閾値ごとの変化がほぼなく，今回の実験では閾値による選別の効果がなかったことが確認できる．

開発データで人手評価との相関が高い平滑化手法の代表として平滑化手法 2 に関して，複数参照応答を利用した BLEU(BLEU-multi) と Δ BLEU による疑似応答数毎の人手評価との Spearam の順位相関係数 ρ の推移を Figure 5 に示す．BLEU-multi と Δ BLEU で両者について，開発データとテストデータでの人手評価との相関の推移について乖離が存在することが確認できる．

テストデータで人手評価との相関が高い平滑化手法の代表として平滑化手法 5 に関して，同様に人手評価との Spearam の順位相関係数 ρ の推移を Figure 6 に示す．BLEU-multi と Δ BLEU で両者について，開発データとテストデータでの人手評価との相関の推移について平滑化手法 2 に比べて乖離が少ないことが確認できる．

この 2 つの平滑化手法 (2, 5) について Δ BLEU による評価での推移をまとめたものを Figure 7 に示す．以上の考察から，平滑化手法 2 では特定の発話ごとに最適な疑似応答サイズが存在し，データに大きく依存しているのに対し，平滑化手法 5 では平滑化手法 2 よりは疑似応答サイズごとの変動が小さく，発話に対する最適な疑似応答サイズが別のデータから見積もりやすいのではないかと推測される．

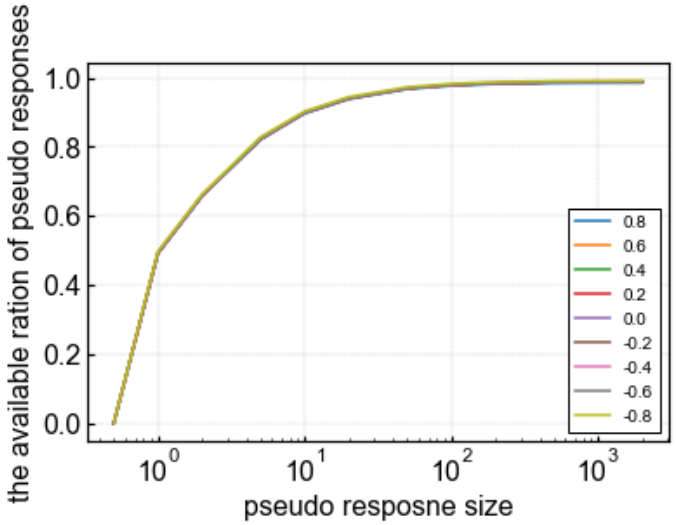

Figure 4　最大疑似応答数に対する各閾値による選別後の疑似応答の利用割合

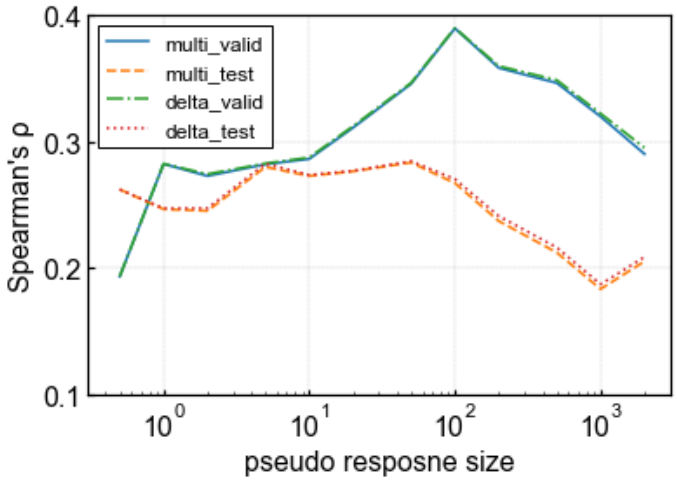

Figure 5　平滑化手法 2 を利用した場合の複数参照による BLEU と Δ BLEU での開発データ・テストデータでの疑似応答数毎の人手評価との Spearman の順位相関係数 ρ の推移

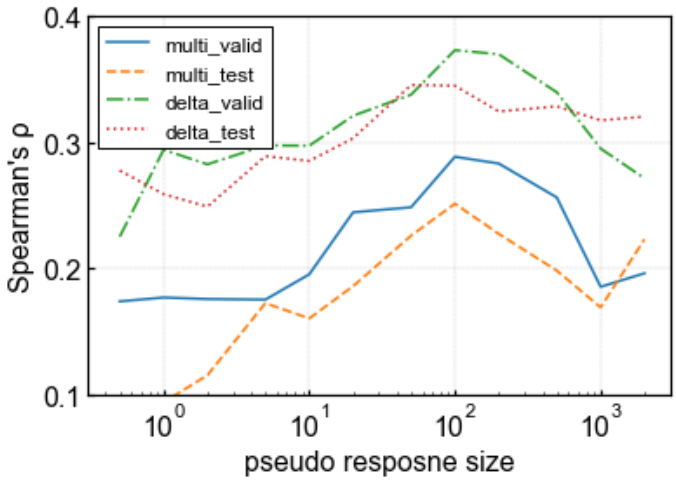

Figure 6　平滑化手法 5 を利用した場合の複数参照による BLEU と Δ BLEU での開発データ・テストデータでの疑似応答数毎の人手評価との Spearman の順位相関係数 ρ の推移

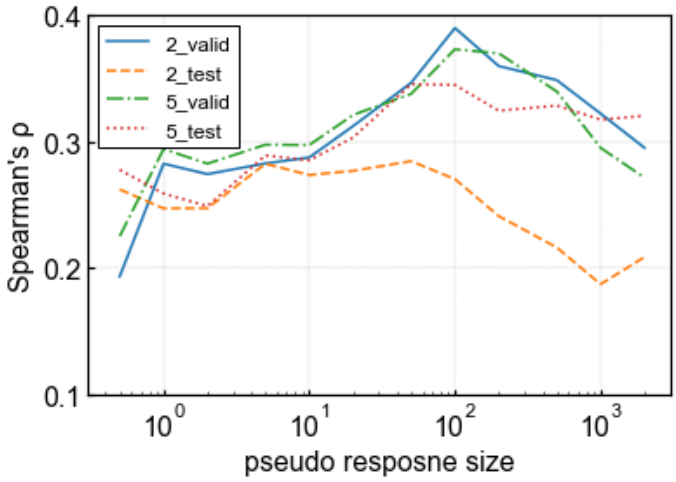

Figure 7　Δ BLEU による評価に関して平滑化手法 2 や 5 を利用した際の開発データ・テストデータでの疑似応答数毎の人手評価との Spearman の順位相関係数 ρ の推移

第 6 章

結論

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，妥当性付き複数参照応答を獲得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，ΔBLEU との比較により，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．また提案手法を単一参照応答に基づく評価手法へ適用し，妥当性付き複数参照応答に基づく評価を行い，結果として単一参照応答のみに基づく評価手法や妥当性を利用しない複数参照応答に基づく評価手法と比較して，人手評価との相関が向上することを確認した．さらに，妥当性付き複数参照応答を利用した評価手法を，雑談対話システムの評価手法 RUBER [9] と組み合わせることで，既存の組み合わせによる評価と比較して，人手評価との相関に関して向上することを確認した．

第 7 章

今後の課題

7.1 提案手法の問題点

今後の課題として，以下の点が挙げられる．

- 提案手法での疑似応答への妥当性付与が分類器の学習が複数応答を持つ発話という Twiiter コーパス等の一部のコーパスでしか十分に収集できないデータに依存していること
- 疑似応答数やその妥当性の閾値など，複数のハイパーパラメタが存在するため，低コストではあるが開発データとして人手評価を行った対話データが必要であること
- 発話によって応答の多様性が変化しうるのに対し，発話に対する疑似応答数が一定

この理由として，一つ目は，オープンドメインな雑談対話システムを評価するためには，この制約があることは大きなペナルティとなっていることは明確である．次に二つ目は，考察でもあったように，必ずしも開発データで調整することが望ましいわけではないこと，また少量とはいえ評価手法のために評価データが必要な状況は望ましいとは言えないことが原因である．最後に三つ目は，考察での異なるの平滑化手法を比較した際に，一部の手法は最適なハイパーパラメタがデータ（発話）に大きく依存している可能性がある．そのため発話に応じて適切なパラメータが動的に選べることで，人手評価との相関が上がる見込みがありうるということである．以上の理由からこれを解決する手法として，特定の条件を課さない発話-応答ペアから学習可能な，ハ

イパーパラメタへの依存の少ない，発話ごとに動的な疑似応答収集手法と妥当性付与手法が求められる．

7.2　提案手法の結果に基づく発展的研究課題

提案手法による妥当性付き疑似応答が，常にどの発話についても利用可能であった場合，一つの応答先として多様性を考慮した，複数応答を持つ発話により学習を行う雑談対話システム [37,38] の学習データの拡張として利用することで，このタスクの発展につながるのではないかと考えられる．

また，雑談対話のサブタスクにはパーソナリティの考慮 [39] や文脈の考慮 [2] といった応答多様性とは共存しにくいタスクが存在する．例えば長い会話のやりとりがあったとしても，常にその会話のやりとりの中で多くの人が応答を行うということはほとんどなく，むしろ特定の個人のみがやり取りをしていることがほとんどだと考えられる．このため，提案手法により将来的に文脈を考慮した応答の拡張が行うことが可能であれば，多大な発展が見込めるのではないかと考えられる．

謝辞

本稿を執筆するに辺り，多くの方に感謝したいと思います．まず初めに指導教官である吉永准教授には論文の執筆の際には休日の遅くまで付き合っていただくこともあり，この研究に関しても多くのご助言をいただき深く感謝いたします．豊田正史教授には，日々収集なさっている会話データありがたくも利用させていただき，また時折この研究に関するアドバイスを伺うことができ非常に参考になりました．喜連川教授は時折しかお会いしませんでしたが，会うたびに研究において何が大事なのかを改めて考えさせられる言葉を頂き，励みになりました研究室の先輩や同期，後輩の方々とは，それぞれの意見を持った活発な議論が行え，多くの刺激を受けることが出来，多くを学べました．その他，秘書・経理の方々は何かとサポートして頂くことがあり，大変助かりました．皆さまのお陰でこの 2 年間，充実した研究生活を送れたことを改めて感謝いたします．

2020 年 1 月 30 日

参考文献

[1] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Association for the Advancement of Artificial Intelligence*, pp. 3776–3783, 2016.

[3] Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. Modeling Situations in Neural Chat Bots. In *Proceedings of ACL 2017, Student Research Workshop*, pp. 120–127, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[5] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[6] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 445–450, Beijing, China, July 2015. Association for Computational Linguistics.

[8] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 196–205, Denver, Colorado, May 2015. Association for Computational Linguistics.

[9] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. In *AAAI Conference on Artificial Intelligence*, 2018.

[10] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, pp. 109–126. National Institute of Standards and Technology (NIST), 1994.

[11] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle

Pineau, Aaron Courville, and Yoshua Bengio. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Association for the Advancement of Artificial Intelligence*, pp. 3295–3301, 2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.

[13] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 285–294, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[14] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 172–180, Los Angeles, California, 2010. Association for Computational Linguistics.

[15] Hochreiter Sepp and Schmidhuber Jurgen. LONG SHORT-TERM MEMORY. *Neural Computation 9(8)*, pp. 1735–1780, 1997.

[16] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[17] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE*

Transactions on Signal Processing, Vol. 45, No. 11, pp. 2673–2681, November 1997.

[18] Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, Barcelona, Spain, July 2004.

[19] Jianfeng Gao and Xiaodong He. Training MRF-Based Phrase Translation Models using Gradient Ascent. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 450–459, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[20] Boxing Chen and Colin Cherry. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[21] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1116–1126, Vancouver, Canada, 2017. Association for Computational Linguistics.

[22] Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 379–391, Stockholm, Sweden, September 2019. Association for Computational Linguistics.

[23] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[24] Gabriel Forgues and Joelle Pineau. Bootstrapping Dialog Systems with Word Embeddings. In *Nips, Modern Machine Learning and Natural Language Processing Workshop*, 2014.

[25] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics.

[26] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, 2018. Association for Computational Linguistics.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations*, 2015.

[28] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap

between Human and Machine Translation. In *CoRR*, September 2016.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[30] Ryuichiro Higashinaka, Noriaki Kawamae, Kugatsu Sadamitsu, Yasuhiro Minami, Toyomi Meguro, Kohji Dohsaka, and Hirohito Inagaki. Building a conversational model from two-tweets. In *2011 IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 330–335, December 2011.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, October 2019.

[33] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, Vol. 70, No. 4, pp. 213–220, 1968.

[34] J. L. Fleiss and M. Davies. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology*, Vol. 115, No. 6, pp. 841–845, June 1982.

[35] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agree-

ment for Categorical Data. *Biometrics*, Vol. 33, No. 1, p. 159, 1977.

[36] J. P. Guilford. *Fundamental Statistics in Psychology and Education*. Fundamental Statistics in Psychology and Education. McGraw-Hill, New York, NY, US, 1942.

[37] Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3826–3835, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. Towards a Neural Conversation Model with Diversity Net Using Determinantal Point Processes. In *AAAI Conference on Artificial Intelligence*, pp. 5932–5939, 2018.

[39] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.

発表文献

査読無し国内会議

1. 蔦侑磨，吉永直樹，豊田正史，応答多様性を考慮した雑談対話評価のための疑似応答獲得．2020 年度 人工知能学会全国大会（第 34 回），熊本, 2020（予定）

研究会

1. 蔦侑磨，吉永直樹，豊田正史，応答多様性を考慮した雑談対話評価のための疑似応答獲得．NLP 若手の会 第 14 回シンポジウム (YANS2019), 北海道, 2019.

2. 蔦侑磨，吉永直樹，豊田正史，疑似応答を用いた雑談対話システムの自動評価．第 6 回自然言語処理シンポジウム / 第 243 回自然言語処理研究発表会，東京，2019.

3. 蔦侑磨，吉永直樹，豊田正史，疑似応答を用いた雑談対話システムの自動評価，東京大学 音声・言語・コミュニケーション研究会 2019，東京，2019.