Twitter コーパスに基づく雑談対話システムにおける多様性の獲得

村田　憲俊†　　酒井　哲也†

† 早稲田大学基幹理工学部情報理工学科　〒 169–8555 東京都新宿区大久保 3-4-1
E-mail: †muratacnttto@suou.waseda.jp, ††tetsuyasakai@acm.org

あらまし　Sequence-to-sequence（Seq2Seq）[1] 生成型ニューラルネットモデルは，雑談対話のタスクで優れた結果を出している．しかし，現在 Seq2Seq は汎用的で意味のない返答をする傾向にあり，対話者の体験を損ないやすくなっている．Twitter のツイートなどノイズの多い文をコーパスとして使用した場合などは，この傾向が顕著である．本研究では，システムの返答の多様性を獲得するために，訓練中の損失関数に着目した手法の提案を行う．また，Twitter の日本語のツイートをコーパスとした，日本語雑談対話システムを作成し，評価を行う．実験の結果，多様性を測るために使用した評価指標 distinct-N のもとで，提案システムは提案手法を使用しない baseline を上回った．また，翻訳の評価指標 BLEU による評価のもとでも，提案システムは baseline を平均的に上回り，前者に対する差は統計的に有意であった（$p \approx 0.0000$）．

キーワード　対話システム，Seq2Seq, テキストマイニング

1 はじめに

Seq2Seq [1] の生成型ニューラルネットモデルは，雑談対話のタスクで優れた結果を出している．しかし元来 Seq2Seq は機械翻訳で用いるために提案されたモデルである．このモデルでは，モデルが入力文に対する正しい出力文を生成する確率について最大化することを目標としている．その目標を達成するために最尤推定を目的関数に使用することが多い．そのため現在 Seq2Seq は雑談対話のタスクで使用する際，汎用的で意味のない返答（e.g.，それはそう思う。）をする傾向にあり，対話者の体験を損ないやすくなっている．また，使用するコーパスにおいて未知語が多い場合や，口語のような文法が正確でない場合など，ノイズの多いときに，この傾向は顕著に現れる．

これに伴い，多様な返答をさせるための研究が増えてきている [2] [3] [4] [5] [6]．これらの手法はモデルを作成するためのそれぞれの工程で，より多様な返答を得るための工夫がされている．ここでは，機械学習のモデルを作成する工程を前処理，訓練，推論で分けて考える．

前処理の段階で Liu らはコーパスの個々の対話について重要度が異なると考え，統計的な手法を用い重み付けを行った [6]．訓練の段階で，Li らは相互情報量の理論に着想を得て，最尤推定の目的関数を再構築し直した [2]．しかし，この手法を用いた際，文法上正しくない返答を返しやすくなった．推論の段階で，Li らはデコーダの出力の決定によく用いられる beam search に注目し，多様な返答を返すように，返答候補の re-rank を行った [3]．各工程で様々な研究がされているが，現在訓練時の目的関数やそれに対して使用される損失関数を再構築する有効な手法が確立されていない．よって，訓練時に適用できる返答の多様性を獲得するための手法が求められる．

また，日本語で Seq2Seq を用いたモデルを作る際には異なる課題が存在する．一般的に，人対人の対話データは非常にプライベートなデータであることが多い．そのことから，大量のオープンドメインな対話データは公開されにくい．現在，データが潤沢に公開されていない言語で Seq2Seq を使用する際に Twitter などの SNS を利用して，擬似的に対話データを作成することが行われている．しかし，多くの場合において SNS を使用した対話データはノイズが多く，このデータをコーパスとしてモデルを作成した場合，汎用的で意味のない返答をしやすくなる．よって，ノイズが多いデータを用いた際に多様な返答をするための有効な手法が求められる．

本研究では，ノイズが多い条件下で，システムが返答の多様性を獲得するために，訓練中の損失関数に着目した手法の提案を行う．また，Twitter の日本語のツイートをコーパスとして，提案手法を使用した日本語雑談対話システムを作成し，評価を行う．実験の結果，多様性を測るために使用した評価指標 distinct-N のもとで，提案システムは提案手法を使用しない baseline を上回った．また，翻訳の評価指標 BLEU による評価のもとでも，提案システムは baseline を平均的に上回り，前者に対する差は統計的に有意であった（$p \approx 0.0000$）．

2 従来研究

本節では，本研究に関連する研究について説明をする．まず，本研究で使用している，Seq2Seq を対話データで使用できるように拡張した手法について紹介をする．次に，Seq2Seq における返答が多様性を獲得するために提案された現在までの手法を紹介する．

2.1 Seq2Seq の対話モデル

2.1.1 Seq2Seq の言語モデル

Seq2Seq は再帰ニューラルネットワーク（Recurrent Neural Network; RNN）[7] である．文（sentence） のような可変長の文字列をシーケンスデータとして扱う．可変長のシーケンス

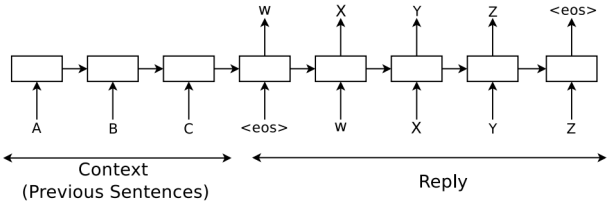

図 1　ニューラルネットワーク対話モデル（文献 [8] より引用）

データを入力とし，可変長のシーケンスデータを出力することができるニューラルネットワークモデルである．まず，Seq2Seq を言語モデルとして利用するための処理を行う．入出力ベクトルの次元数（語彙数）と処理系内部での次元数を合わせるために，以下の式が導入される．

$$x_t = W_{x'x} x'_t \tag{1}$$

$$y_t = \mathrm{softmax}(W_{hy} h_t) \tag{2}$$

入力シーケンス（単語列）は $\{x'_1, x'_2, ..., x'_{T_i}\}$ で，x'_* は語彙数と同じ次元数のベクトルである．語彙を表現するために，該当するインデックスの次元の値のみが 1 となり，それ以外の値は 0 になるような one-hot ベクトルになっている．出力シーケンスは $\{y_1, y_2, ..., y_{T_o}\}$ で，y_* は，時系列における各時刻の予測単語の確率分布ベクトルである．softmax はソフトマックス関数であり，出力値を正規化し，確率分布として扱うことができるようにする．

上記の言語モデルの学習では，入力シーケンスは単語列に該当し，$\{x'_1(=\langle bos\rangle), x'_2, ..., x'_{T_i}\}$ を入力することにより，それぞれの次の単語が何であるかを順番に予測していく．$\langle bos\rangle$（beginning of sentence）は文頭を表現するためのシンボル単語である．言語モデルの場合，期待される出力単語列は $\{y'_1 = x'_2, y'_2 = x'_3, ..., y'_{T_i} = \langle eos\rangle\}$ となる．これは，入力単語の次の単語が何であるかを予測していることになる．$\langle eos\rangle$（end of sentence）は文末を表現するためのシンボル単語である．本研究ではこれに加え，語彙の中に存在しない単語のシンボル単語として $\langle unk\rangle$（unknown）を導入している．$\langle bos\rangle$ や $\langle eos\rangle$，$\langle unk\rangle$ といった特殊なシンボル単語は使用される語彙として事前に one-hot ベクトルの次元として組み込まれる．出力単語列 y'_* は予測出力単語の確率分布を表すベクトルのため，期待される出力単語列の出力確率は以下のような同時確率となる．

$$p(y'_1, y'_2, ..., y'_T | x'_1, x'_2, ..., x'_T) = \prod_{t=1}^{T} p(y'_t | c_{t-1}, h_{t-1}, x'_t) \tag{3}$$

2.1.2　Seq2Seq の対話モデルへの拡張

Seq2Seq [1] を元に，Vinyals らは，対話データを使えるように一部拡張をし，ニューラルネットワークを用いた対話モデルとして提案した [8]．以下に，その提案手法での拡張部分の詳細を述べる．

対話モデルの構造の概要は 2.1.1 の言語モデルと同様である．対話モデルの構造について図 1 に示す．

拡張部分は，主に以下の 2 点である．

I　発話者の切り替えを示すためのシンボル文字の導入

II　学習時に予測する出力単語列を返答文のみとする

I は，対話には発言者と返答者がいるが，入力単語列で発話者の切り替えを表現するために発言終了のシンボル単語 $\langle eos\rangle$ を導入する．入力単語列中にこのシンボル単語が出現した場合，それは発話者の切り替えを意味し，返答者が発話を開始する．

II は，学習時に予測する出力単語列は返答文のみであり，発言文に関しては無視をするということである．相手の発言文を $\{x_{i*}\}$ とし，それに対するシステムの返答文を $\{x_{o*}\}$ とする．入力単語列は以上のことから $\{x_{i*}, \langle eos\rangle, x_{o*}, \langle eos\rangle\}$ となる．これに対して，対話モデルでは $\{\square, ..., \square, x_{o*}, \langle eos\rangle\}$ を予測すればよく，$\square$ に関しては評価をしない．

2.2　多様性を獲得するための手法

ここでは，Seq2Seq が多様性を獲得するために提案されている代表的な手法について紹介する．Liu らはコーパスの個々の対話について学習時の重要度が異なると考え，統計的な手法を用い重み付けを行った [6]．この手法では，Seq2Seq の返答の多様性が損なわれる原因は，コーパスのデータを等価として最尤推定を行うことであると考えている．そのため，2 つの観点でコーパスに重み付けを行い，最尤推定後に多様性を損なわないようにしている．1 つ目は，ある返答がコーパス中に多数出現する場合は，その返答は汎用的なものであると考える．また，2 つ目はシステムの返答は短すぎたり，長過ぎることを避けるべきであると考える．つまり，コーパス中に存在する返答の内，繰り返し使用されず，平均的な長さであるものが重要であるということである．重みは以下の式で定義される．

$$w(y|x, \mathbb{R}, \mathbb{C}) = \frac{\Phi(y)}{\max_{r\in\mathbb{R}} \Phi(r)} \tag{4}$$

ここで，x が入力発話．y がそれに対する返答．$\mathbb{R}$ が x に対する返答として，コーパス $\mathbb{C}$ に存在するものの集合である．また，個々の返答の重みは以下のように推測される．

$$\Phi(y) = \alpha\mathcal{E}(y) + \beta\mathcal{F}(y) \tag{5}$$

$\mathcal{E}$ と $\mathcal{F}$ は各観点の重みを推測するための関数である．α と β はハイパーパラメータであり，使用者が任意に設定を行う．

$\mathcal{E}$ は返答のコーパス中での出現頻度に関する重み付けの関数である．式は以下のようになる．

$$\mathcal{E}(y) = \mathrm{e}^{-af(y)} \tag{6}$$

$$f(y) = \max\{0, \mathrm{Count}(D(y, y_j) \geqq \tau) - b\} \quad \forall j \in |\mathbb{C}| \tag{7}$$

Count は集合の個数を数え上げる関数．D は コーパスの各返答間の距離を計算する関数である．この研究では，n-grams での一致度を用いた．a はスケール因子，τ は閾値，b はバイアスに当たるハイパーパラメータである．$\mathcal{F}$ は返答のコーパス中での長さに関する重み付けの関数である．式は以下のようになる．

$$\mathcal{F}(y) = \mathrm{e}^{-c||y|-|\hat{y}||} \tag{8}$$

$$|\hat{y}| = \frac{1}{|\mathbb{C}|}\sum_{r\in\mathbb{R}} |r| \tag{9}$$

c はスケール因子に当たるハイパーパラメータである．

以上の式を用いて，コーパス中の各返答の重み付けを行い，その重みを学習時のパラメータ更新に反映をさせる．

3 提案手法

本節では，ノイズが多い条件下で，システムが返答の多様性を獲得するために，訓練中の損失関数に着目した手法の提案を行う．この提案手法には 2 つの観点が存在する．それぞれの観点に対応して 3.1 では，未知語へのペナルティに関する手法，3.2 では文頭の多様化に関する手法を提案する．

3.1 未知語へのペナルティ

ノイズが多い条件下では，モデルは汎用的で多様性が損なわれた返答をする傾向にある．SNS をコーパスとする場合のノイズには，特に未知語の問題が存在する．SNS ではユーザーのコミュニケーションのために返答内でユーザー名などがよく使用される．個人名に当たるものは，コーパス全体で見たときに個別には使用頻度が少なく，情報も得にくいため，未知語としてまとまって扱う場合が多い．しかし，SNS では返答の対象者を特定するために，コーパス内で個別での使用頻度は少ないが，ユーザー名という集合で見た際には，使用頻度が非常に高い．そのため，コーパス内で未知語が多くなりやすい傾向にある．また，日本語の場合は単語分割の問題が存在し，SNS などの文の形態素解析は精度が落ちやすい．そのため，単語が正確に分割できないことから，実際に使用している単語のユニーク数より，使用単語のユニーク数が多くなる傾向にある．実装をする際，メモリの都合で使用単語数を制限する場合が多いので，このことも未知語を多くする原因となっている．

以上のことを受けて，未知語への対策をするために，以下の損失を導入する．

$$\ell_{unk}(x, \mathbb{B}, \theta) = \sum_{y \in \mathbb{B}} \sum_{w \in y} \log p(w = \langle unk \rangle | x; \theta) \tag{10}$$

$\mathbb{B}$ は訓練中のバッチ全体である．つまり，バッチごとに見た際に，すべての返答の各単語で未知語を意味する $\langle unk \rangle$ を使用する確率が高い場合，損失が大きくなるようにした式になる．

3.2 文頭の多様化

Seq2Seq は最尤推定を行った結果，コーパス内に多く現れる表現を使いやすくなり，どのような場面でも使いやすいような汎用的な表現を多用する傾向にある．これに対して，従来研究では，コーパス内に多く現れる文について，重み付けを軽くすることにより対処していた [6]．本研究では，訓練の段階で使用することのできる手法として，損失関数の拡張に関する手法を提案する．この手法では，Seq2Seq のデコーダ部分に着目をする．デコーダは推論時には $\langle bos \rangle$ を最初の入力とし，各単語の出力について自分の生成した前の単語を入力として出力を生成する．そのことから，Seq2Seq はデコーダがネットワークとして表現力の高いものである場合，エンコーダでの入力文の解釈を無視し，自らの出力に依存する傾向にある．ここでは，その傾向を活用する手法を提案する．

デコーダの出力は $\langle bos \rangle$ を入力した際の出力から始まる．デコーダはここでの出力を次の入力としてその後の文を生成するため，この手法は入力文の最初の単語を多様化させるためのものである．ここで，文頭だけでなく文全体で多様化するための拡張を行わないのは，訓練時に損失として，文全体の多様化を目指してしまうと，Li らの研究 [2] で確認されたような，文法上正しくない返答をしてしまう危険があるためである．以下に導入した損失を示す．

$$\ell_{head}(x, \mathbb{B}, \theta) = \sum_{\mathbb{V}} \left(\sum_{y \in \mathbb{B}} p(y_0 | x; \theta) \right)^2 \tag{11}$$

$\mathbb{V}$ はすべての語彙を意味する．バッチごとに見た際に，文頭単語の確率のノルムが大きい場合，損失が大きくなるようにした式になる．つまり，文頭の単語の使用頻度が偏ることを避ける方向の損失関数である．

上記の 2 つの観点の損失を元の最尤推定をするための損失関数と合わせると次のような式になる．

$$loss(x, \hat{y}, \mathbb{B}, \theta) = - \sum_{y \in \mathbb{B}} \log p(y = \hat{y} | x; \theta) + \alpha \ell_{unk} + \beta \ell_{head} \tag{12}$$

$\hat{y}$ は正解データ，α, β はスケール因子に当たるハイパーパラメータである．

式 (12) を損失関数として，Seq2Seq における返答の多様性の獲得を行う．

4 評価・実験

提案手法の性能を評価するために比較実験を行った．比較実験では，同一のニューラルネットワーク構造を持つ Seq2Seq について，提案手法を使用した場合と使用しなかった場合のモデルで比較を行った．4.1 では評価に使用したデータセットの説明，4.2 では評価を行う上で設定をした各種数値等の説明，4.3 では実験結果の説明を行う．

4.1 データセット

対話の学習データとして，Twitter のツイートを利用した．Twitter は現在，日本国内で幅広く利用されており，数多くのジャンルの会話が行われている．よって，Twitter のツイートで構成されるコーパスは雑談対話システム作成し得ると考えられる．また，ツイートは日本語の文法的な乱れやネットスラングが含まれていることが多いため，Seq2Seq の返答の多様性が損なわれやすいコーパスである．

本研究では任意のツイートと，そのツイートに対するリプライというペアを対話データとして扱った．学習データ全体は 2014/01/01 から 2014/12/31 までに取得された 1,000,000 (1M) ツイートを対象としており，500,000 ペアの対話のデータを用意した．

評価を行うために，上記の学習データを $train, test, val$ の 3 つのデータセットに分割した．分割をする際，今回のデータは

1 年間のデータを取得しているため，日付に対してランダムに抽出を行った．これは，例えば年始の挨拶のようなツイートが $train, test, val$ のいずれかに偏るような状況を避けるためである．$train$ はモデルの学習のために用いる．$test$ はモデルの評価のために用いる．val はモデルが学習をする際に，モデルが $train$ に対して過学習を行わないようにするために，一定バッチ数ごとにモデルを検証し，学習を制御するために用いる．各データセットのサイズについて，表 1 に示す．

表 1　データセットのサイズ

種類	割合	サイズ
ALL	100	500,000
train	76.5	344,250
val	13.5	60,750
test	10	50,000

4.2 実験設定

以下，本研究での実験条件詳細について述べる．

まず，入力シーケンスである $\{x_{w*}\}$ を得るための日本語テキストデータの形態素解析については $Janome$ [9] を使用した．文字列長 T については，テキストチャットは一般的に文字列長が短い傾向にあると考えられる．よって，長いものについてはチャット以外の用途で用いられると仮定し，$T = 30$ とした．シンボル文字 $\langle pad \rangle$ は後ろ詰めで調整をした．

実験に使用したモデルについて説明を行う．式 (12) におけるハイパーパラメータは $\alpha = 1.0$，$\beta = 0.5$ とした．Seq2Seq における embedding layer には，日本語の Wikipedia [10] の記事を元に学習した fastText [11] を使用した．encoder には，Bidirectional LSTM [12] を 2 層重ねたものを使用し，decoder では，LSTM を highway connections [13] したものを使用した．また，これらの実装には AllenNLP [14] を用いた．推論の段階では，beam search を使用した．その際の width は 5 とした．学習の際，100 batch ごとに val データセットを用いて，学習中のモデルの検証を行った．そのとき，val データセットへの損失が上昇した場合，$train$ に対して過学習をしていると考え，学習を止めるようにした．

4.3 実験結果の算出手法

本研究では，各手法を比較するために Li らが多様性を測るために用いた distinct-N [2] と BLEU [15] [16] を使用する．

distinct-N は以下の式により計算される．本研究では，$N \in \{1, 2\}$ とする．

$$\text{distinct-}N = \frac{\#distinct\ N\text{-}grams\ of\ generated\ tokens}{total\ number\ of\ generated\ tokens} \quad (13)$$

この式はシステムが生成した全ての返答の N-gram でのユニーク数と総単語数の比となる．

また，BLEU は以下の式により計算される．

$$BLEU = BP * PREC \quad (14)$$

$$BP = \exp(\min(0, 1 - \frac{SBML}{SYSL})) \quad (15)$$

$$SBML = \sum_s BML(s) \quad (16)$$

$$= \sum_s \arg\min_{len(s^*)} |len(s) - len(s^*)| \quad (17)$$

$$SYSL = \sum_s len(s) \quad (18)$$

$$len(s) = \text{文 } s \text{ の長さ} \quad (19)$$

$$PREC = \exp(\frac{1}{2} \sum_{N \in \{1,2\}} lnPrec_N) \quad (20)$$

$$Prec_N = \frac{\sum_s \sum_{e \in gram_N(s)} Clip(c, s)}{\sum_s \sum_{e \in gram_N(s)} C(e, s)} \quad (21)$$

$$Clip(e, s) = \min(\max_{s^*} C(e, s^*), C(e, s)) \quad (22)$$

$$C(e, s) = s \text{ の中に含まれている単語 } e \text{ の個数} \quad (23)$$

今回の場合はそれぞれ，s はシステムの返答文，s^* はテストデータの返答文，$gram_N(s)$ は文 s から得られる N-gram の集合，e は文 s を形態素解析をした単語である．また，$N \in \{1, 2\}$ とした．すなわちユニグラムとバイグラムのみを考慮した．

4.4 実験結果

Twitter のデータから作成した対話のデータセットを用いて，提案手法の比較を行った．評価には $test$ データセットを使用した．baseline として，提案手法を使用していない Seq2Seq を用いる．

まず，各システムが生成した返答を用いて distinct-N を算出する．$test$ での distinct-N について表 2 に示す．

表 2　各システムの distinct-N の比較

システムの種類	distinct-N
baseline	0.000489
proposed model	0.001564

distinct-N において，提案手法は baseline より優れた数値を示した．これは提案手法が baseline よりも多様な単語を使用して，返答を生成していることを意味する．

次に，baseline と提案手法に関して，$test$ に対してのシステムの返答と，実際のユーザーの返答を用いて BLEU を算出する．$test$ での BLEU の平均について表 3 に示す．

表 3　各システムの BLEU の比較

システムの種類	Mean BLEU
baseline	0.002932
proposed model	0.008416

システム間の平均の差について，上記の差が統計的な有意な差であるかを確かめるために t 検定を行った．$\alpha = 0.05$ としたとき，この結果の p 値は $p \approx 0.0000$ であった．よって，システム間の平均に差がないという帰無仮説は棄却され，統計的に有意な差であることが確認された．

表 4　各システムの出力例

発言文	フォローありがとうございます !!
正解	⟨*usr*⟩ サンキュー
baseline	いえいえ
proposed model	こちらこそよろしく
発言文	おはよう、みんな！きょうも、がんばるです！
正解	おはよー ⟨*usr*⟩…今日も早いわねぇ
baseline	フォロバしまし
proposed model	おはよう
発言文	おやすみなさいませ
正解	おやすみなさい…
baseline	⟨*unk*⟩
proposed model	おやすみなさい
発言文	デスティニーやりたいけどお高いから買えない…
正解	余裕あるときに買うと良いと思うの
baseline	⟨*unk*⟩ さんです
proposed model	大敵です

各システムが出力した例を表 4 に示す．

出力例では，⟨*bos*⟩ と ⟨*eos*⟩ は省略している．また，実際のデータでは ⟨*usr*⟩ に Twitter 上のユーザー名と推測されるものが入っている．

上記の出力例でも一部確認ができるように，baseline と比較して，提案手法は ⟨*unk*⟩ を使用する頻度が少なくなっていた．beam search における，確率の高い返答候補の中で，出力単語に ⟨*unk*⟩ を含む割合は，baseline では 0.8174 に対し，提案手法は 0.5864 であった．これは，システムができる限り ⟨*unk*⟩ を使用せずに返答を生成するように学習を行っていることを意味する．

5 考　　察

本研究の実験では，Twitter のツイートを対話のデータセットとして扱ったが，BLEU の値が提案手法も含め非常に低い数値となった．この原因は，ツイートとそれに対するリプライだけでは，任意の発話に対しての適切な返答を学習するためには情報が少ないことであると考えられる．一般的に SNS で行われる会話等のコミュニケーションをする際には，何らかのドメインを共有していること多いが，今回のデータでは文脈に相当する情報がなかったため，適切な返答をすることが困難な会話も多かった．そのため，挨拶や Twitter のフォローなどの機能に関するやりとり等，返答がある程度定型化しているものについては比較的適切な返答が得られたが，それ以外の会話については不適切な返答も多かった．

返答の多様性の獲得について，未知語へのペナルティをすることにより ⟨*unk*⟩ が使用されることを抑制し，文頭の多様化により返答を結果的に多様化させることに成功した．しかし，現状では文頭の多様化の重みを強くし過ぎると，文意の通らない文を生成することが多くなるので，ハイパーパラメータの調整が重要となっている．これはシステム作成時の計算コストが大きくなる原因の一つであり，今後の改善が求められる点である．

本研究では対話システムの返答の妥当性を評価をするために，評価指標として BLEU を用いた．しかし，非タスク指向の対話システムを評価する上では BLEU は最適な評価指標であるとは言い難い [17]．よって，適切な評価指標自体を検討する必要がある．

6 ま と め

本研究では，システムの返答の多様性を獲得するために，訓練中の損失関数に着目した手法の提案を行った．また，Twitter の日本語のツイートをコーパスとした，日本語雑談対話システムを作成し，評価を行った．多様性を測るために使用した指標機械 distinct-N のもとで，提案システムは提案手法を使用しない baseline を上回った．また，翻訳の評価指標 BLEU による評価のもとでも，提案システムは baseline を平均的に上回り，前者に対する差は統計的に有意であった（$p \approx 0.0000$）．今回提案した手法は関連研究と組み合わせて使用することができるため，今後は複数の手法を組み合わせて，多様性があり妥当な返答が得られるシステムについて研究を行う必要がある．

文　　献

[1] Sutskever I, Vinyals O, Le. Quoc V “Sequence to Sequence Learning with Neural Networks,” NIPS 2014, pp. 3104–3112, 2014.

[2] Li J, Galley M, Brockett C, Gao J, Dolan B “A Diversity-Promoting Objective Function for Neural Conversation Models” Proceedings of NAACL-HLT 2016, pp. 110-119, 2016.

[3] Li J, Monroe W, Jurafsky D “A Simple, Fast Diverse Decoding Algorithm for Neural Generation” arXiv preprint arXiv:1611.08562, 2016.

[4] Lison P, Bibauw S “Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models,” Proceedings of the SIGDIAL 2017 Conference, pp. 384–394, 2017.

[5] Mou L, Song Y, Yan R, Li G, Zhang L, Jin Z “Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation” Proceedings of the International Conference on Computational Linguis-tics (COLING), 2017.

[6] Liu Y, Bi V, Gao J, Liu X, Yao J, Shi S “Towards Less Generic Responses in Neural Conversation Models: A Statistical Re-weighting Method” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2769-2774, 2018.

[7] Michael C. Mozer “A Focused Backpropagation Algorithm for Temporal Pattern Recognition” Complex Systems 3(4), pp. 349–381, 1989.

[8] Vinyals O, Le Quoc V, “A Neural Conversational Model,” ICML 2015 Deep Learning Wrokshop, 2015.

[9] 打田 智子, “janome”, 2015, https://github.com/mocobeta/janome, （参照 2019/01/05）

[10] “Wikipedia”, https://dumps.wikimedia.org/jawiki/, （参照 2019/01/05）

[11] Joulin A, Grave E, Bojanowski P, Mikolov T, “Bag of Tricks for Efficient Text Classification,” Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431, 2017.

[12] Mike S, Kuldip K. P, “Bidirectional Recurrent Neural Net-

works," IEEE TRANSACTIONS ON SIGNAL PROCESSING, Volume. 45, NO. 11, pp. 2673–2681, 1997.

[13] Zilly J, Srivastava R, Koutnik J, Schmidhuber J, "Recurrent Highway Networks," arXiv preprint arXiv:1607.03474, 2016.

[14] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.H., Peters, M., Schmitz, M., Zettlemoyer, L.S. "A Deep Semantic Natural Language Processing Platform." , 2017.

[15] Papineni K, Roukos S, Ward T, Zhu W, "BLEU: a Method for Automatic Evaluation of Machine Translation," the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318, 2002.

[16] 酒井 哲也, "情報アクセス評価方法論 検索エンジンの進歩のために," コロナ社, 2015.

[17] C.W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation," Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2122–2132, 2016.