E-018

Encoder-Decoder ニューラル対話モデルにおける 単語埋め込みベクトルのノルムの調査
Investigation of norms of word embedding vectors in encoder-decoder neural conversational model

冨岡 愛也† 岸田 優輝† 加藤 恒夫†
Manaya Tomioka Yuki Kishida Tsuneo Kato

1. はじめに

近年，AI アシスタントやチャットボットの普及に伴い，非タスク指向型対話の応答生成技術が注目されている．対話コーパスを学習データとしてモデルベースで対話応答を生成する技術として，Encoder-Decoder 再帰型ニューラルネットワーク(Recurrent Neural Network, RNN)の応用が広く検討されている[1,2]．Encoder-Decoder RNN は可変長の入出力に対応し，Sequence to sequence で学習することができる[3]．また，入力トークンと出力トークンの対応付けを明確にする注意機構[4,5]を追加することで，精度改善が図られている．

Encoder-Decoder RNN において，入力文を構成するトークン（形態素）はまずエンコーダの Embedding 層で単語埋め込みベクトルと呼ばれる実数ベクトルに変換された後 LSTM や GRU[6]などの再帰層に入力される．出力文のトークンも再帰的にデコーダに入力され Embedding 層で単語埋め込みベクトルに変換された後，同様に再帰層に入力される．エンコーダとデコーダの Embedding 処理は相似であるが，Encoder-Decoder モデルの成り立ちとしては，デコーダはニューラル言語モデル(Recurrent Language Model, RLM)[7]であり，その隠れ層の初期化手法としてエンコーダが開発された[8]ため，両者の Embedding 処理の働きは必ずしも同じではない．

そこで，本稿では Encoder-Decoder ニューラル対話モデルにおけるエンコーダとデコーダの働きを理解するために，両者の Embedding 層を調査する．単語埋め込みベクトルは多次元実数ベクトルであるが，後段の処理で線形変換され，活性化関数に入力されるため，ベクトルのノルムは統計的にモデル出力への影響の大きさを反映していると考えられる．

具体的には，語彙サイズ約 3 万の注意機構付き Encoder-Decoder ニューラル対話モデルを学習し，エンコーダ，デコーダそれぞれの埋め込みベクトルのノルムについて，学習データ中の出現頻度との関係，次単語予測能力の指標として同単語を条件とする Bi-gram のエントロピーとの関係を調べた．

† 同志社大学大学院理工学研究科
Graduate school of science and engineering,
Doshisha University

2. 注意機構付き Encoder-Decoder モデル

2.1 モデル構造

エンコーダは Embedding 層 と GRU 層の 2 層，デコーダは Embedding 層 と GRU 層 と Softmax 関数を持つ全結合層の 3 層と注意機構から成る．図 1 にモデル構造を示す．

入力文を $S = s_1, s_2, \ldots s_m$ これに対する応答文を $T = t_1, t_2, \ldots t_n$ とする．入力トークンの埋め込みベクトルを $\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2} \ldots, \boldsymbol{x}_{s_m}$，$j$ 番目の出力確率を $\boldsymbol{y}_j$ とする．j 番目の出力トークンは式(1)により決定される．

$$t_j = \underset{w}{\operatorname{argmax}}\{\boldsymbol{y}_j(w)\} \quad (1)$$

エンコーダの Embedding 層は，入力トークン s_i を多次元実数ベクトルである埋め込みベクトル $\boldsymbol{x}_{s_i}$ に変換する．

$$\boldsymbol{x}_{s_i} = \mathrm{Embedding}^e(s_i) \quad (2)$$

続く GRU は，各時刻の Embedding 層の出力 $\boldsymbol{x}_{s_i}$ を受け取り，エンコーダの隠れ層ベクトル $\boldsymbol{h}^e{}_i$ を更新し，ベクトル $\boldsymbol{e}_i$ を出力する．

$$(\boldsymbol{e}_i, \mathbf{h}^e{}_i) = \mathrm{GRU}^e(\boldsymbol{x}_{s_i}, \boldsymbol{h}^e{}_{i-1}) \quad (3)$$

ただし，$\boldsymbol{h}^e{}_i$ は零ベクトルで初期化する．

デコーダは，入力文 S に対して，グリーディに最尤のトークンを出力する．隠れ層をエンコーダの最終の隠れ層ベクトルで初期化（$\boldsymbol{h}^d{}_0 = \boldsymbol{h}^e{}_m$）し，最初の入力として"<start>"を与える．デコーダの Embedding 層は，入力トークン t_j を多次元実数ベクトル $\boldsymbol{x}_{t_j}$ に変換する．

$$\boldsymbol{x}_{t_j} = \mathrm{Embedding}^d(t_j) \quad (4)$$

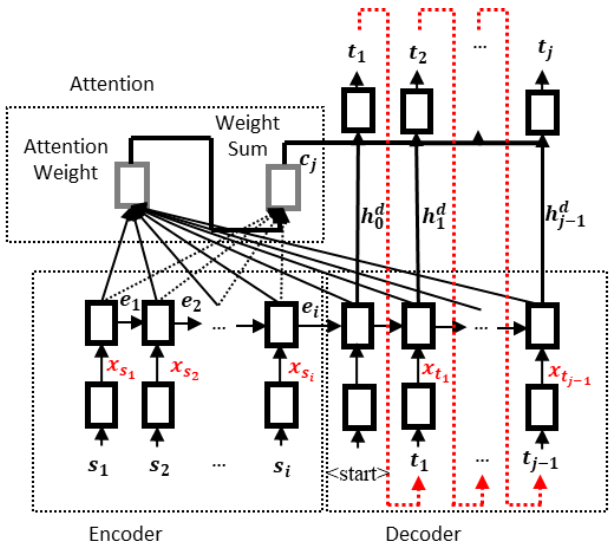

図 1 注意機構付き Encoder-Decoder モデル

注意機構は出力トークンt_jを入力トークン$s_{i,}$と対応付けるため，エンコーダ出力ベクトル系列$e_{1,}e_2, \ldots e_m$をスタックした全隠れ層ベクトル e_{all} と j-1 番目のデコーダ隠れ層ベクトル $h^d{}_{j-1}$から式 (5) により注意重みベクトル a_jを計算する．式 (6)によりエンコーダ出力ベクトルを注意重み$a_j(i)$で重み付け和をとることでコンテキストベクトル c_j を出力する．

$$a_j = \text{Softmax}\{v^T \tanh(W_1 e_{all} + W_2 h^d{}_{j-1})\} \quad (5)$$

$$c_j = \sum_{i=1}^{m} (a_j(i)\, e_i) \quad (6)$$

W_1, W_2, v は重みパラメータである．

注意機構が出力するコンテキストベクトル c_j と直前のデコーダ出力 t_j を入力として出力確率 y_j を得る．y_j の次元数は出力形態素の種類数である．j番目のy_jの確率分布はデコーダ出力ベクトルd_jに基づき求め，Softmax 関数を用いて正規化する．

$$(d_j, h^d{}_j) = \text{GRU}^d(x_{t_j}, c_j) \quad (7)$$

$$y_j = \text{Softmax}(W_y d_j + b_y) \quad (8)$$

ここで，W_yは重み行列，b_yは重みベクトルである．

モデルの学習は，教師付き学習で行う．クロスエントロピー誤差を損失関数とし，ミニバッチごとに損失関数を最小化するように勾配降下法によりモデルパラメータを更新する．デコーダの入力は teacher forcing により常に正解を与える．

2.2 雑談対話コーパスを用いたモデルの学習

対話コーパスとして，2 名の 20 代女性のペルソナを設定してクラウドソーシングによって収集した計 168 万発話からなる仮想雑談対話コーパスを利用した．2 名のペルソナを仮に A と B とする．A の発話から始まり，B の応答が 3 種類示される．B の発話それぞれに対して A が 3 種類の応答を返す．これを繰り返すことで各分岐 が 3 の木が形成される．木の深さは最大 10 である．最初の A の発話は 50 種類あり，50 個の木が形成される．コーパス全体で合計 168 万文の入力応答文対が得られる．

今回は B の入力文と A の応答文の対である約 123 万文をを用いる．以下の前処理を行う．

1. 日本語形態素検索エンジン MeCab[9]で形態素に分割.
2. 「？」を除く記号や句読点を除去.
3. 出現頻度が 1 回の形態素を<oov>に置換.
4. 文の先頭と末尾に<start>，<end>を付与.
5. 入力文の形態素長が一定数になるように<end>の後ろに<pad>詰め.

これを学習・検証・評価用に 18：1：1 に分割する．分割後のデータ数を表 1 に示す．

語彙は入力文用と応答文用で共通化し，コーパス中に 2 度以上出現した 37785 語とした．Embedding 層，隠れ層の次元数をそれぞれ 128 次元，256 次元とし，Embedding 層は 0 で初期化し，GRU の各ゲートの重みは grolot の一様分布で初期化した．バッチサイズは 128，活性化関数は sigmoid 関数，最適化は Adam とする．そして対話モデルをエポック 20 で学習を行った．

共通の入力文に対するコーパス中の応答文と学習エポック 10，20 それぞれモデルが出力する応答文の例を表 2 に示す．多くの先行研究で報告されているとおり，学習が進むにつれて，”そうですよねー”などのありきたりな応答が増える傾向にある．

表 1　18:1:1 に分割したデータセットのサイズ

データセット	入力応答文対	総形態素数
学習	1.1M	21M
検証	62k	1.1M
評価	62k	1.1M

表 2　共通の入力文に対する応答文の例

入力文	ビールなんかも飲みますか？
コーパス中の応答文	ビールは苦手ですね美味しさがわからなくて
エポック 10	ビールは好きですよ大好きです
エポック 20	ビールは好きですよ
入力文	そうですねできれば今より近いところが良いですよね
コーパス中の応答文	満員電車に長時間乗るのは嫌ですね
エポック 10	通勤時間は短い方がいいですね
エポック 20	そうですよねー
入力文	休みが何日かあるとたまにはいいかもですね
コーパス中の応答文	連休はあまりないので欲しいです
エポック 10	休みの日は時間が足りないですよね
エポック 20	そうですよねーそう思います

3. 単語埋め込みベクトルのノルムの調査

3.1 ノルム調査の方法

Embedding 層から埋め込みベクトルのノルムを求め，学習セットにおける出現頻度との関係を調べる．また，埋め込みベクトルのノルムと学習セットにおけるその形態素を条件とする Bi-gram 確率のエントロピー（以下エントロピーとする） との関係を調べる．これをエポック 10，20 それぞれ求める．

エンコーダの埋め込みベクトルのノルムについては学習セットの入力文の形態素の出現頻度，デコーダの埋め込みベクトルのノルムについては学習セットの応答文の形態素の出現頻度との関係を求める．

同様に，次形態素の予測のしやすさとの関係を調べるために，エンコーダの埋め込みベクトルのノルムと学習セットにおける入力文の形態素のエントロピー，および，デコーダの埋め込みベクトルのノルムと学習セットにおける応答文の形態素のエントロピーとの関係を求める．ただしコーパス中の出現頻度が小さい形態素の埋め込みベクトルは適切に学習されない可能性があることから，出現頻度を一定数以上に限定した場合の相関も測定した．

多数ある形態素の中でも，助詞と助動詞に着目した．機能語である助詞と助動詞は種類が少なく，出現頻度が多く，

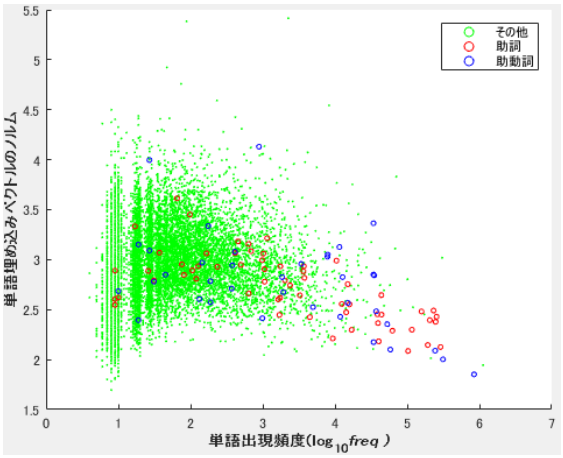

図 2　エポック 10：エンコーダの埋め込みベクトルのノルムと出現頻度の散布図

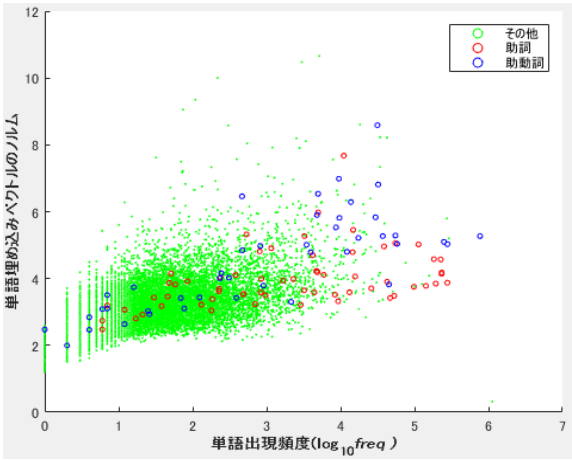

図 3　エポック 10：デコーダの埋め込みベクトルのノルムと出現頻度の散布図

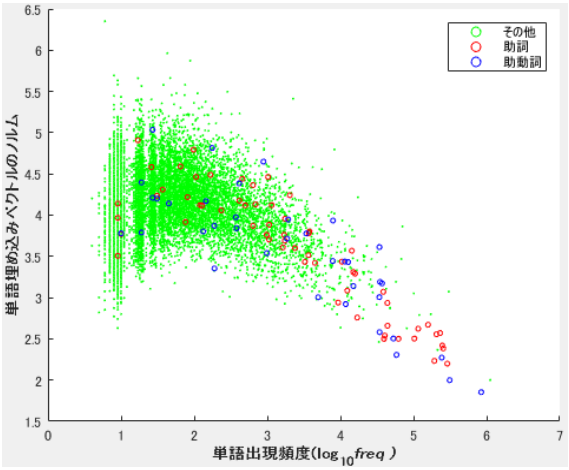

図 4　エポック 20：エンコーダの埋め込みベクトルのノルムと出現頻度の散布図

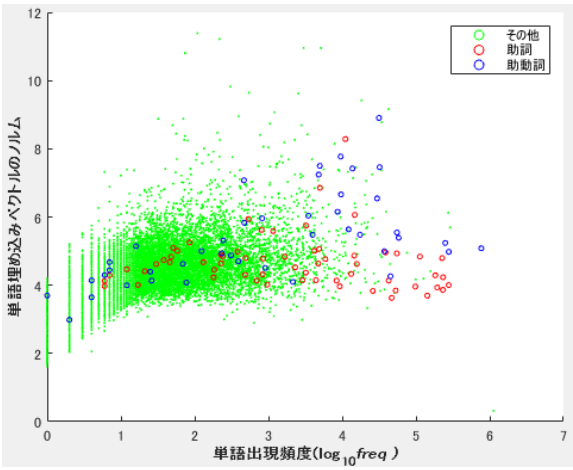

図 5　エポック 20：デコーダの埋め込みベクトルのノルムと出現頻度の散布図

特定の内容を運ばない．助詞は後続の形態素の種類が多いのに対して，助動詞は後続の形態素の種類が少ない．

また，エンコーダ，デコーダそれぞれの隠れ層には，埋め込みベクトルとともに，直前の時刻の隠れ層が影響するが，両者の影響を比較するために，各層のベクトルのノルムの相関を求める．具体的には，評価セットの入力文に対して応答文を生成し，エンコーダ，デコーダそれぞれにおいて GRU 隠れ層ベクトルと埋め込みベクトルのノルムの相関，GRU 隠れ層ベクトルと直前の隠れ層ベクトルのノルムの相関を求める．

3.2 評価指標

i 番目の形態素に対する埋め込みベクトルのノルム$||x_i||$は式（9）で求める．ここでg_{i1}，g_{i2}，…，g_{in}は埋め込みベクトルの各要素を示す．

$$||x_i|| = \sqrt{{g_{i1}}^2 + {g_{i2}}^2 + \cdots + {g_{in}}^2} \tag{9}$$

エンコーダ，デコーダにおける i 番目の形態素w_iの出現頻度とは，それぞれ学習セットの入力文に現れるw_iの出現回数，応答文に現れるw_iの出現回数とする．出現頻度は常用対数をとって表現した．

次形態素の予測のしやすさとの関係を調べるため，学習セットにおいて特定の形態素 i を条件とする Bi-gram 確率のエントロピーを式（10）で求める．

$$Ent(i) = -\sum_{k \in w} \mathrm{P}(k|i)\log_2 \mathrm{P}(k|i) \tag{10}$$

ここで，本研究で扱う対話コーパス中の全語彙wに含まれる形態素を i,k とする．対話コーパス中の学習セットから求めた Bi-gram 確率を$\mathrm{P}(k|i)$とする

3.3 実験結果

図 2，3 にエポック 10，図 4，5 にエポック 20 としたときのエンコーダ，デコーダそれぞれの埋め込みベクトルのノルムと出現頻度との関係を示す．縦軸は埋め込みベクトルのノルム，横軸は学習セットにおける出現頻度を表す．表 3，4 にエンコーダとデコーダにおいて図 2〜5 の散布図から得られた埋め込みベクトルのノルムと出現頻度の相関係数をまとめる．エンコーダにおいて全形態素を対象にすると相関はみられないが，出現頻度を限定し，低出現頻度の形態素を除外すると負の相関がみられた．エポック 20 はエ

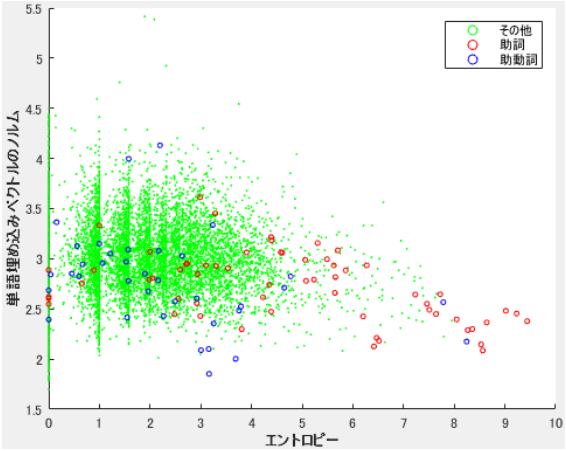

図 6 エポック 10: エンコーダの埋め込みベクトルのノルムと学習セットにおけるエントロピーの散布図

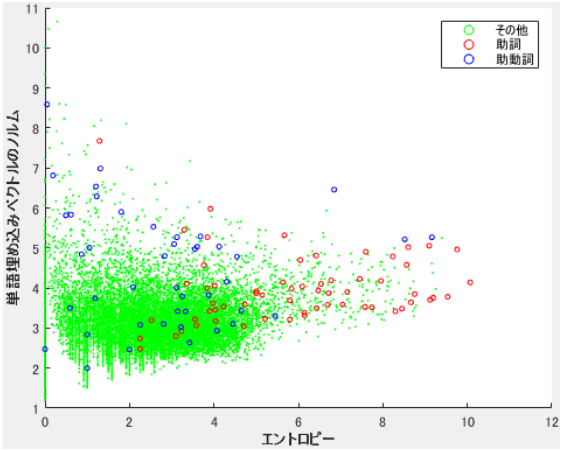

図 7 エポック 10: デコーダの埋め込みベクトルのノルムと学習セットにおけるエントロピーの散布図

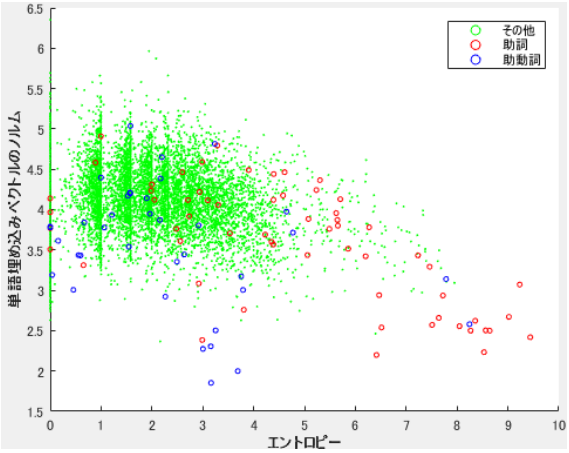

図 8 エポック 20: エンコーダの埋め込みベクトルのノルムと学習セットにおけるエントロピーの散布図

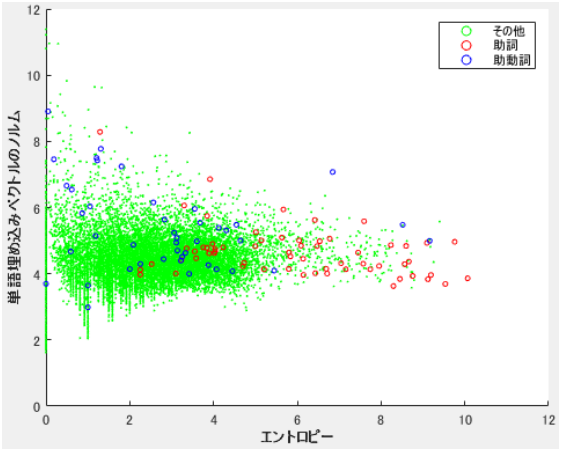

図 9 エポック 20: デコーダの埋め込みベクトルのノルムと学習セットにおけるエントロピーの散布図

ポック 10 に比べて負の相関が強くなった．一方，デコーダにおいては逆に正の相関がみられた．全形態素を対象にすると強い正の相関が現れ，出現回数を限定すると相関が弱くなった．エンコーダにおける助詞や助動詞は，比較的埋め込みベクトルのノルムが小さかった．

図 6，7 にエポック 10，図 8，9 にエポック 20 におけるエンコーダ，デコーダそれぞれの埋め込みベクトルのノルムと学習セットにおけるエントロピーとの関係を示す．縦軸は埋め込みベクトルのノルム，横軸はエントロピーを表す．表 5，6 に図 6〜9 の散布図から得られた埋め込みベクトルのノルムとエントロピーの相関係数をまとめる．ここで，出現頻度が僅かな形態素のエントロピーは必ず低いため除外し，出現頻度が 10 以上（$freq \geqq 10$）のものに限定した．表 5 よりエンコーダではいずれのエポックでも出現頻度を限定しても相関は弱かった．出現頻度との相関よりも弱い．表 6 より，デコーダの埋め込みベクトルのノルムとエントロピーとの相関は出現頻度を限定すると強い負の相関が現れ，エポック 20 の方が負の相関が強かった．

表 7 に評価セットから応答文を生成する際のエンコーダ，デコーダにおける GRU 隠れ層ベクトルと埋め込みベクトルのノルム，および GRU 隠れ層ベクトルと直前の時刻の GRU 隠れ層ベクトルのノルムの相関を示す．エンコーダの GRU 隠れ層ベクトルのノルムは埋め込みベクトルのノルムと直前の時刻の GRU 隠れ層ベクトルのノルムで強い正の相関がみられる．一方，デコーダは GRU 隠れ層ベクトルのノルムと埋め込みベクトルのノルムで強い正の相関がみられるが，直前の時刻の GRU 隠れ層ベクトルのノルムとの相関は比較的小さい．

4. 考察

エポック 10，20 で学習を行ったエンコーダにおける埋め込みベクトルのノルムと出現頻度に関して出現頻度を限定するほど強い負の相関がみられる．しかし，埋め込みベクトルのノルムと Bi-gram のエントロピーとは明確な相関が現れなかった．また，助詞や助動詞は内容語に比べて埋め込みベクトルのノルムが全般的に小さかった．これらを総合するとエンコーダにおける埋め込みベクトルのノルムは応答文生成に与える意味的な強さを反映していると考えられる．

エポック 10，20 で学習を行ったデコーダにおける埋め込みベクトルのノルムと出現頻度との間には，全形態素を対象とすると強い正の相関がみられたが，出現頻度を限定すると相関は弱かった．そこで詳細を調査したところ，出現

表 3 エンコーダの埋め込みベクトルの
ノルムと出現頻度との相関係数

	全形態素 (17084 語)	*freq*>=10 (10764 語)	*freq*>=100 (3369 語)
エポック 10	0.156	-0.083	-0.276
エポック 20	-0.082	-0.369	-0.619

表 4 デコーダの埋め込みベクトルの
ノルムと出現頻度との相関係数

	全形態素 (35203 語)	*freq*>=10 (12310 語)	*freq*>=100 (3402 語)
エポック 10	0.752	0.445	0.347
エポック 20	0.748	0.301	0.149

頻度が 10 未満（*freq*<10）の低頻度形態素が全体の約 2/3 を占めており，その中でも出現頻度が 1〜3 のものが特に多数を占めた．出現頻度が 1〜3 の低頻度形態素についてはそのノルムと出現頻度に強い正の相関があるため，全形態素の相関係数も結果的に高くなった．エポック 20 で学習し出現頻度を 100 以上（*freq*≧100）に限定すると，埋め込みベクトルのノルムはエントロピーとの間に強い負の相関がみられ，単語出現頻度との相関よりも強くなる．また，表 7 より，埋め込みベクトルのノルムはデコーダの GRU 隠れ層ベクトルのノルムに強い影響を与えている．これらのことから，デコーダにおける埋め込みベクトルのノルムは次の出力形態素を予測する言語モデルとしての強さを反映していると考えられる．

5. おわりに

本稿では日本語の大規模な仮想雑談対話コーパスを用いて学習した Encoder-Decoder ニューラル対話モデルにおけるエンコーダとデコーダの埋め込みベクトルのノルムを調査した． Embedding 層から形態素の埋め込みベクトルを抽出し，そのノルムを測った．そして学習セットにおける出現頻度，および学習セットにおいてその形態素を条件としたときの Bi-gram 確率のエントロピーとの関係性を分析した．

エンコーダでは埋め込みベクトルのノルムと出現頻度の間に負の相関がみられた．助詞や助動詞は比較的埋め込みベクトルのノルムは小さいものが多かった．エンコーダにおける埋め込みベクトルのノルムは応答文生成に与える意味的な強さを反映していると考えられる．

デコーダでは埋め込みベクトルのノルムと出現頻度に関してエンコーダとは反対に強い正の相関がみられた．しかし，出現頻度を限定すると，埋め込みベクトルのノルムと出現頻度との相関よりも，埋め込みベクトルのノルムとエントロピーとの間の相関が強くなった．さらに，デコーダの GRU 隠れ層への入力と出力の相関を調べることにより，デコーダの GRU 隠れ層ベクトルのノルムは埋め込みベクトルのノルムの影響をより強く受けていることを確認した．これらの結果からデコーダにおける埋め込みベクトルのノルムは次の出力形態素を予測する言語モデルとしての強さを反映していると考えられる．

表 5 エンコーダの埋め込みベクトルのノルムと
学習用セットにおけるエントロピーとの相関係数

	freq>=10 (10764 語)	*freq*>=100 (3369 語)
エポック 10	-0.045	-0.182
エポック 20	-0.222	-0.246

表 6 デコーダの埋め込みベクトルのノルムと
学習用セットにおけるエントロピーとの相関係数

	freq>=10 (12310 語)	*freq*>=100 (3402 語)
エポック 10	-0.145	-0.288
エポック 20	-0.215	-0.471

表 7 GRU 隠れ層ベクトルのノルムと埋め込みベクトル
のノルム，直前の隠れ層ベクトルのノルム
それぞれとの相関係数

	埋め込み ベクトル	直前の時刻の 隠れ層ベクトル
Encoder	0.642	0.517
Decoder	0.511	0.274

参考文献

[1] O. Vinyals and Q. Le, “A neural conversation model” ICML(2015).
[2] J. Li, M. Galley, C. Brockett, J. Gao and B. Dolan, “A diversity-promoting objective function for neural conversation models”, In Proc. of NAACL-HLT 2016, pp.110-119 (2016).
[3] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks", In Proc. of NIPS 2014, pp.3104-3112 (2014).
[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, arXiv:1409.0473(2014).
[5] M. T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation”, In Proc. of EMNLP 2015, pp.1412-1421 (2015).
[6] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio “Learning phrase representations using RNN encoder-decoder for statistical machine translation” In Proc. of EMNLP 2014, 1724-1734 (2014).
[7] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, “Recurrent neural network based language model”, INTERSPEECH 2010, pp.1045–1048 (2010).
[8] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao and B. Dolan, “A Neural network approach to context-sensitive generation of conversational responses”, In Proc. of NAACL-HLT 2015, 196-205 (2015).
[9] 工藤拓, 山本薫, 松本 裕治, “Conditional Random Fields を用いた日本語形態素解析” , In proc. of EMNLP, pp.230-237(2004).