一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2019-12(2019-08)

# WordNetを用いた雑談対話システムの汎化性能の向上

麻生　大聖†　　高島　遼一†　　滝口　哲也†　　有木　康雄†

† 神戸大学　〒657–8501 兵庫県神戸市灘区六甲台町 1–1

**あらまし**　日本語 WordNet を用いて，LSTM Encoder-Decoder による雑談対話システムの汎化性能を向上させる手法について検討する。雑談は非常に幅広い話題を扱い，表現も多種多様であるため，汎化性能を向上させることが困難である。例えば，あまり現れないマイナーな単語や表現がユーザから入力されたときに，関係性の低い応答を生成する恐れがある。そこで，ユーザ入力文に含まれる単語の分散表現に，その上位語の分散表現を加算して入力することで，単語を概念的に広く捉えて，学習不足な単語や表現が入力されても適切な応答を生成することを目的とする。日本語 WordNet とは Princeton WordNet と呼ばれる大規模言語データベースに日本語が付与されたものであり，各単語の上位・下位概念および上位・下位語を検索することができる。応答文の入力文との関係性を，日本語 WordNet を用いない場合と比較した。

**キーワード**　雑談対話システム，汎化性能，WordNet，上位概念，上位語，単語の分散表現

# Improvement of Generalization Performance of Non-task-oriented Dialogue System by Use of WordNet

Taisei ASO†, Ryoichi TAKASHIMA†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Kobe University　1–1 Rokkodai-cho, Nada-ku, Kobe-shi, Hyogo, 657–8501 Japan

## 1. はじめに

近年，IoT 化に伴ってテキストチャットや音声による会話型インターフェースが拡大しており，人間とやりとりができる対話システムの研究が盛んに行われている。NTT ドコモ社の『しゃべってコンシェル』や，Apple 社の『Siri』などは，ユーザの質問や要求に対して適切な情報を提供したり，アプリケーションの操作をする一方で，雑談を行うことでユーザをサポートしている。このように雑談という機能は，人間とのやりとりを円滑にする重要な役割を担っている。

Twitter などのソーシャルネットワーキングサービスから大量に対話データを収集し，機械学習を行うことで，ユーザ入力文に対する雑談応答を生成することが可能である。しかし，雑談は非常に幅広い話題を扱い，表現も多種多様である。そのため，汎化性能を向上させることが困難であり，あまり現れないマイナーな単語や表現がユーザから入力されたときに，関係性の低い応答を生成する恐れがある。

本研究では，大規模言語データベースである日本語 WordNet を用いて，それらの問題を抑制することを目的としている。日本語 WordNet を用いることで，指定した単語の上位・下位概念や，それに属する単語などを検索することができる。入力文に含まれる単語をその上位語を加算した単語ベクトルに変換して入力することで，入力文の意味を概念的に広く捉えて，ユーザ入力に含まれる様々な単語や表現に柔軟に対応することが期待できる。応答文の入力文との関係性を，日本語 WordNet を用いない場合と比較した。

## 2. WordNet

Princeton WordNet [1] は，単語が類義関係のセット (Synset) でグループ化された英語の大規模言語データベースである。各 Synset には固有 ID が割当てられており，それぞれが一つの概念に対応している。各単語は一つ以上の Synset に属しており，各 Synset は上位・下位関係などの様々な関係で結ばれている。

日本語 WordNet [2] は，Princeton WordNet の Synset に対応して日本語が付与されており (Fig. 1)，Princeton WordNet に存在しない Synset も含んでいる。収録された Synset 数や単語数，語義数は次のとおりである。

- 57,238 概念 (Synset 数)
- 93,834 単語
- 158,058 語義 (Synset と単語のペア)

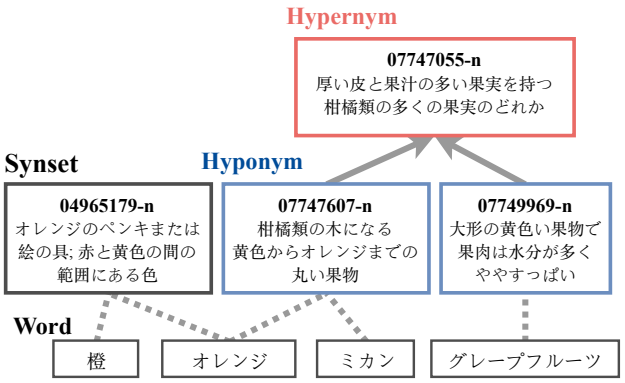


図 1　日本語 WordNet
Fig. 1　Japanese WordNet

## 3.　データセット

### 3.1　Twitter 対話コーパス

ソーシャルネットワーキングサービス『Twitter』におけるツイートとリプライのペアを対話データとして収集した。本研究では対話履歴を考慮しないため，複数回の返信による対話であっても，入力文と応答文のペアに分割した。英数字や顔文字などの特定の文字や，画像・URL などの外部情報を含むペアは除去した。単語数が 4 以上かつ 40 以下の，名詞を含むペアのみを取り出し，句読点や繰り返し表現を正規化し，合計 51 万の対話データを用意した。その中の 50 万を学習データ，1 万を評価データとして使用した。MeCab [3] を用いて形態素解析を行った。

Fig. 2 では使用した対話データに含まれる品詞ごとの単語の種類数を総数で割った値 (Distinct) を比較している。特に名詞が他の品詞と比べて，非常に多様であることがわかる。

### 3.2　Word2Vec 学習用の Wikipedia 記事

本研究では，単語の分散表現に Word2Vec [4]〜[6] を用いた。Word2Vec の学習には，Twitter から収集した学習用対話データに加えて，インターネット百科事典『Wikipedia』の日本語版記事データを用いた。Wikipedia 記事データには Twitter 対話データと同様のフィルタリング処理と正規化処理を施し，合計 3,049,628 文 (381.7MB) になった。

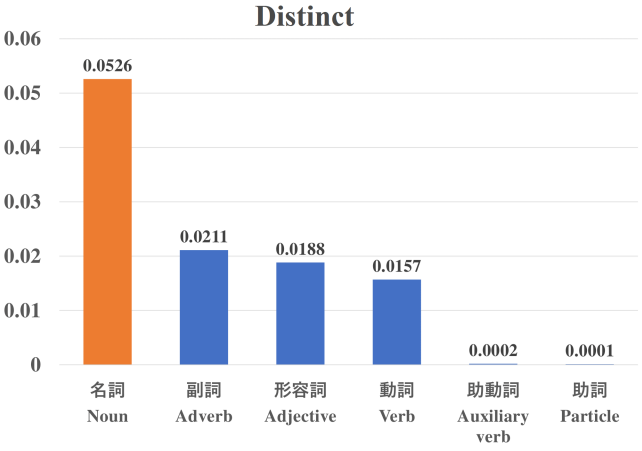


図 2　Twitter 対話コーパス中の各品詞の種類数を総数で割った値
Fig. 2　Distinct of each part of speech in Twitter dialogue corpus

## 4.　研 究 手 法

### 4.1　ベースライン

本研究では，雑談応答生成に Fig. 3 のような LSTM による RNN Encoder-Decoder [7] を用いた。入力系列の時系列を逆転させて，各単語を分散表現 Word2Vec に変換して入力する。

### 4.2　提 案 手 法

日本語 WordNet では，名詞と動詞の上位・下位概念がサポートされている。また Fig. 2 から，名詞が他の品詞と比べて多様であるため，全ての単語や表現を学習することができず，関係性の低い応答や無難な応答を生成する原因の一つとなっていると考えられる。そこで，入力系列に含まれる名詞 $w$ を，式 (1) から式 (6) により $V$ に変換して入力する手法を提案する。

式 (1) では，入力された名詞 $w$ を含む概念が日本語 WordNet に存在するかを確認し，存在しなければ，その名詞の Word2Vec による分散表現を $V$ とする。存在すれば，名詞 $w$ を含む全ての概念 $s$ の概念ベクトル $SV$ を式 (2) により計算し，その単純平均を $V$ とする。提案手法の概略図を示した Fig. 4 では，「サッカー」を含む概念 04167661-n と概念 00478262-n の概念ベクトルの単純平均を，「サッカー」の分散表現としている。

式 (2) では，入力された概念 $s$ に含まれる全ての日本語単語の Word2Vec による分散表現の単純平均と，概念 $s$ の全ての上位概念の概念ベクトルの単純平均を重み付き加算し，再帰的に概念ベクトルを計算する。概念 $s$ に日本語単語が含まれなければ，代わりに概念 $s$ の全ての上位概念の概念ベクトルの単純平均を概念ベクトルとする。Fig. 4 では，概念 00467719-n に日本語単語が含まれないため，代わりにその上位概念 00464651-n の概念ベクトルと置き換えている。

これにより，入力系列の意味を概念的に広く捉えることが期待できる。共通の上位概念をもつ単語同士は，近い分散表現として入力されるため，マイナーな単語にも対応することが期待できる。

この手法には二つのパラメータ *ratio* と *depth* がある。*ratio* は 0 から 1 までの範囲の値であり，*depth* は非負整数である。*ratio* は上位概念の加算比重を示し，$ratio = 0$ のときには同じ概念に含まれる類義語のみを加算することになる。*depth* は加算する上位概念の最大の深さを示す。

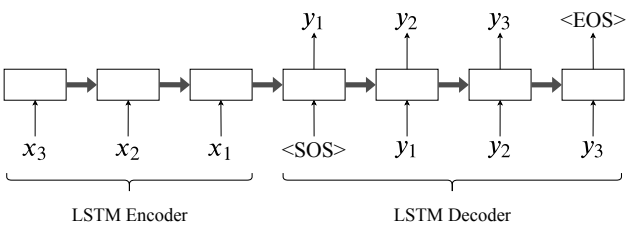


図 3　LSTM Encoder-Decoder ベースラインモデル
Fig. 3　LSTM Encoder-Decoder baseline model

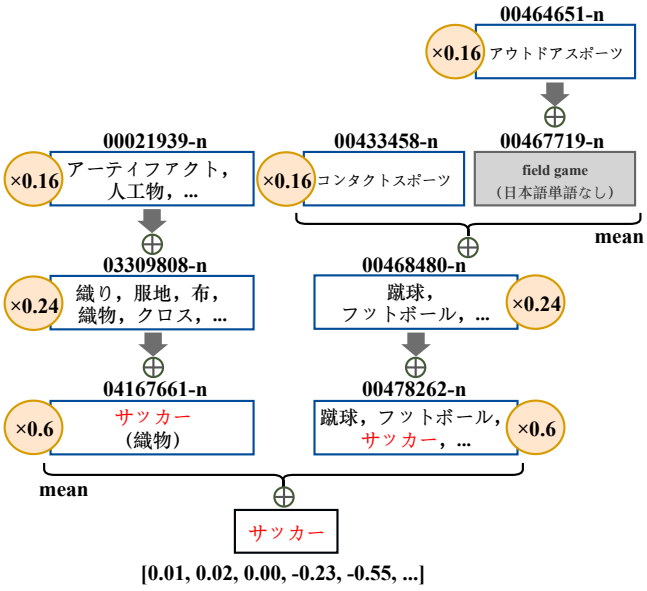


図 4　提案手法の概略図 ($ratio = 0.4, depth = 2$ の例)
Fig. 4　Proposed Method ($ratio = 0.4, depth = 2$)

$$V(w) = \begin{cases} (\text{単語 } w \text{ の } Word2Vec), & |\, W2S(w)\,| = 0 \\ \dfrac{\sum_{s \in W2S(w)} SV(s, depth)}{|W2S(w)|}, & otherwise \end{cases} \quad (1)$$

$$SV(s,d) = \begin{cases} WV(S2W(s)), & |\, S2H(s)\,| = 0 \text{ or } d = 0 \\ \dfrac{\sum_{h \in S2H(s)} SV(h,d)}{|S2H(s)|}, & |\, S2W(s)\,| = 0 \\ (1 - ratio) WV(S2W(s)) + ratio \dfrac{\sum_{h \in S2H(s)} SV(h, d-1)}{|S2H(s)|}, & otherwise \end{cases} \quad (2)$$

$$WV(ws) = \frac{\sum_{w \in ws} (\text{単語 } w \text{ の } Word2Vec)}{|\, ws\,|} \quad (3)$$

$$W2S(w) := (\text{単語 } w \text{ を含む概念の集合}) \quad (4)$$

$$S2W(s) := (\text{概念 } s \text{ に含まれる単語の集合}) \quad (5)$$

$$S2H(s) := (\text{概念 } s \text{ の上位概念の集合}) \quad (6)$$

## 5. 実　　験

### 5.1 実 験 条 件

Word2Vec の学習パラメータは Table 1 のように設定した。LSTM Encoder-Decoder のパラメータは Table 2 のように設定した。提案手法において，上位概念の加算比重 $ratio$ と，加算する上位概念の最大の深さ $depth$ は Table 3 のように設定し，4 種類の $ratio$ を比較した。

表 1　Word2Vec の学習パラメータ
Table 1　Parameters of Word2Vec training

| | |
|---|---|
| 学習モデル | Skip-gram |
| 次元数 | 256 |
| 文脈長 | 5 |
| 単語最低出現数 | 5 |
| 語彙数 | 250,908 |
| 反復回数 | 10 |

表 2　LSTM Encoder-Decoder のパラメータ
Table 2　Parameters of LSTM Encoder-Decoder

| | |
|---|---|
| ユニット数 | 256 |
| 隠れ層数 | 3 |
| 出力語彙数 | 32,302 |
| 最適化手法 | Adam [8] |
| 初期学習率 | 1e-4 |
| ドロップアウト率 | 20% |
| バッチサイズ | 256 |
| 学習エポック数 | 300 |
| ビームサーチ幅 | 15 |

表 3　提案手法のパラメータ
Table 3　Parameters of proposed method

| | |
|---|---|
| $ratio$ | 0.1, 0.2, 0.3, 0.4 |
| $depth$ | 2 |

### 5.2 分散表現の分布の比較

Word2Vec による分散表現の主成分分析を行った。寄与率は第一主成分は 6.09% で，第二主成分は 5.08% であった。Word2Vec による分散表現と提案手法による分散表現を主成分軸に合わせて次元圧縮した。それらの分布を Fig. 5〜9 に示す。提案手法では，「水泳」と「スイミング」などの類義語や，「バドミントン」と「テニス」という同じコート競技の単語同士は，近くに分布していることがわかる。また，Word2Vec では「ダイヴィング」は語彙に存在しないためゼロベクトルとして扱われていたが，提案手法では日本語 WordNet に存在するため類義語である「ダイビング」と同じベクトルになっている。

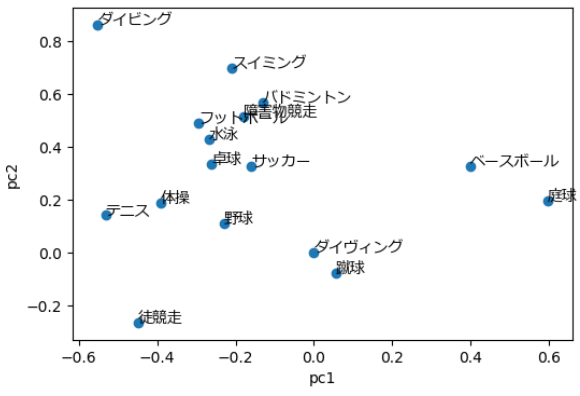


図 5　Word2Vec による分散表現の分布
Fig. 5　PCA of Word2Vec distributed representation

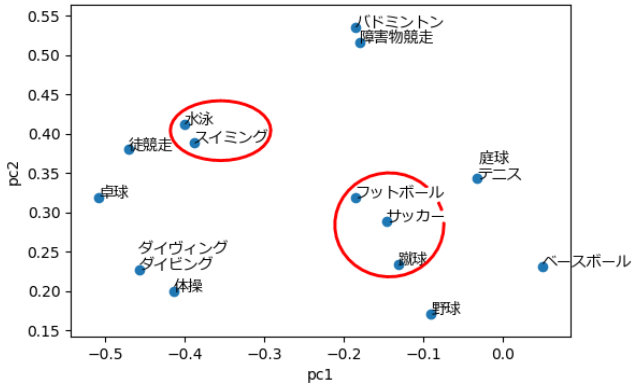


図 6　提案手法による分散表現の分布 ($ratio = 0.1$)
Fig. 6　PCA of proposed distributed representation ($ratio = 0.1$)

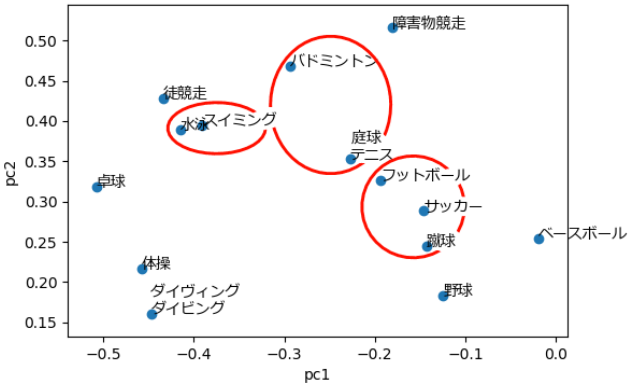


図 8　提案手法による分散表現の分布 ($ratio = 0.3$)
Fig. 8　PCA of proposed distributed representation ($ratio = 0.3$)

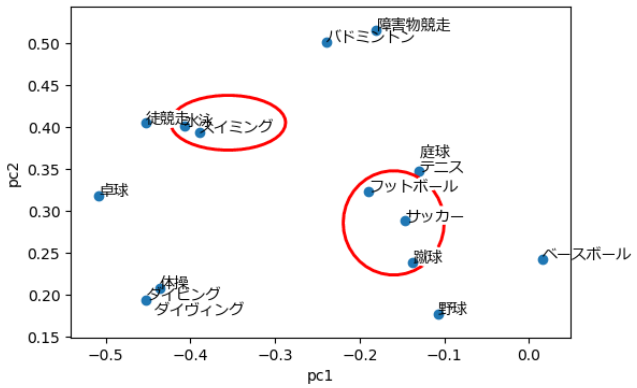


図 7　提案手法による分散表現の分布 ($ratio = 0.2$)
Fig. 7　PCA of proposed distributed representation ($ratio = 0.2$)

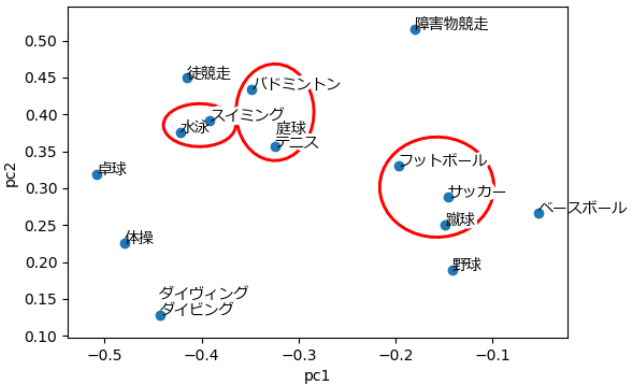


図 9　提案手法による分散表現の分布 ($ratio = 0.4$)
Fig. 9　PCA of proposed distributed representation ($ratio = 0.4$)

### 5.3　実験結果と考察

#### 5.3.1　BLEU による客観評価

Twitter から収集した 1 万対話を評価データとして，各手法により応答文を生成した。Table 4 で各手法による応答文の BLEU [9] を比較している。括弧内の数値はベースラインに対しての BLEU の増加率を表す。

提案手法では 4 種類の *ratio* において，ベースラインよりも BLEU が向上した。BLEU-1 は $ratio = 0.2$ のときに最大となり，ベースラインに対して 0.00224 (1.74%) 向上した。BLEU-2 は $ratio = 0.3$ のときに最大となり，ベースラインに対して 0.001503 (7.55%) 向上した。$ratio = 0.1$ では BLEU はあまり増加せず，また *ratio* の値を大きくし過ぎると逆効果となる傾向がみられた。上位語を加算することで，マイナーな単語であっても意味を概念的に広く捉えて，学習データに多く含まれるメジャーな単語と近い分散表現に変換することができたため，学習が容易になり BLEU が向上したと考えられる。しかし，*ratio* の値を大きくし過ぎると，多くの単語が似た分散表現に変換されて区別ができなくなるため，学習が困難になり BLEU の増加率が減少したと考えられる。

本実験で用いた Twitter 対話コーパスは，文の読点を「、」に正規化し，文の終わりは「。」「！」「？」「!?」のいずれかに正規化した。したがって，これらの文字はほぼ全てのデータに存在し，学習が容易であったため BLEU-1 が BLEU-2 に比べて大きくなったと考えられる。

雑談は非常に幅広い話題を扱う複雑なタスクであり，BLEU 評価と人手評価には差異が生まれることが考えられるため，今後はアンケートによる主観評価も行うことも検討している。

表 4　各手法による BLEU
Table 4　BLEU of each method

| | BLEU-1 | BLEU-2 |
|---|---|---|
| ベースライン | 0.128396 | 0.019919 |
| 提案手法 ($ratio = 0.1$) | 0.129724 (+1.03%) | 0.020449 (+2.66%) |
| 提案手法 ($ratio = 0.2$) | **0.130636 (+1.74%)** | 0.021222 (+6.54%) |
| 提案手法 ($ratio = 0.3$) | 0.129314 (+0.71%) | **0.021422 (+7.55%)** |
| 提案手法 ($ratio = 0.4$) | 0.128854 (+0.36%) | 0.020936 (+5.11%) |

#### 5.3.2 応答文の比較

各手法によって生成した応答文の例を Table 5 に示す。頻繁に出現する単語や表現が入力されたときは，どちらの手法でも適切な応答を生成できることが多かった。しかし，Word2Vec の語彙に含まれない単語や，学習データに多く出現しないマイナーな単語が入力されたときに，ベースラインでは無難な応答や関係性の低い応答を生成することが多かったが，上位語を加算した単語ベクトルに変換する提案手法ではそれが抑制された。

Word2Vec は，同じ文脈に出現する単語同士は似た意味を持つという分布仮説に基づいているが，近い概念の単語同士が近い単語ベクトルになるとは限らない。対して提案手法では，近い概念の単語同士はまとまるように変換されている。例えば，「パンケーキ」は比較的出現回数の少ない単語であったが，上位語に「ケーキ」や「オーブンで焼かれた食品」のような単語をもつため，共通の上位語をもつ「パン」や「洋菓子」などの出現回数の多い単語と近い単語ベクトルに変換されて扱われた。

通常，雑談においては，入力文のトピックを逸脱しないように応答文が返されるものと考えられる。例えば，Twitter において，入力文が食べ物についての内容であるときは，応答文は「美味しそう」などの表現を含むことが多い。上位語を加算することにより，そのようなトピック情報を付与することができたため，BLEU が向上したのではないかと考える。

しかし，いずれの手法においても，意味的に破綻しているような応答文がみられた。対話破綻抑制の機構を取り入れることで，さらなる汎化性能の向上が期待できると考えている。

## 6. おわりに

本研究では，入力文に含まれる名詞の分散表現を，上位語を加算したものに変換して，雑談応答生成を行った。ベースラインよりも BLEU が向上し，マイナーな単語にも対応しやすくなった。しかし，応答文全体を見ると，入力文に対して意味的に破綻することがあり，破綻抑制が課題である。

今後は，WordNet などの言語知識を用いて，語義・概念に対しても単語ベクトルと同じ次元の分散表現を導出する AutoExtend [10] のような手法を用いた雑談応答生成を検討している。

**謝辞** 本研究の一部は，JSPS 科研費 JP17K00236, JP17H01995 の助成を受けたものである。

### 文　献

[1] Princeton University "About WordNet." WordNet. Princeton University. 2010, http://wordnet.princeton.edu

[2] Francis Bond *et al.*, “Enhancing the Japanese WordNet,” *ALR7 Proc. the 7th Workshop on Asian Language Resources*, pp. 1–8, *Association for Computational Linguistics*. pp. 1–8, 2009.

[3] Taku Kudo, “Mecab: Yet another part-of-speech and morphological analyzer,” http://mecab.sourceforge.net/, 2005.

[4] Tomas Mikolov *et al.*, “Linguistic regularities incontinuous space word representation,” *Proc. of NAACL-HLT 2013*, pp. 746–751, 2013.

[5] Tomas Mikolov *et al.*, “Efficient estimationof word representations in vector space,” arXiv:1301.3781, 2013.

[6] Tomas Mikolov *et al.*, “Distributed Representations of Words and Phrases and their Compositionality,” In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[7] Ilya Sutskever *et al.*, “Sequence to Sequence Learning with Neural Networks,” In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[8] Diederik Kingma and Jimmy Ba, “Adam: A method for stochastic optimiza-tion,” In *The International Conference on Learning Representations (ICLR)*, 2015.

[9] George Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics,” *Proc. of the Second International Conference on Human Language Technology Research 2002 (HLT ’02)*, pp. 138–145, 2002.

[10] Sascha Rothe and Hinrich Schutze, “AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes,” *Proc. of ACL 2015*, pp. 1793–1803.

表 5　応答文の例

Table 5　Generation examples

| | |
|---|---|
| 入力文 | よろしくお願いします！呼びタメ大歓迎なので、気軽に絡んでください。 |
| 正解文 | わかりました。早速ですが、なんてお呼びしたら良いですか？ |
| ベースライン | ありがとうございます！私も呼びタメ大歓迎なので気軽に絡んでください。 |
| 提案手法 ($ratio = 0.1$) | こちらも呼びタメ大歓迎なので仲良くしてください。なんて呼んだらいいかな？ |
| 提案手法 ($ratio = 0.2$) | 私も呼びタメ大歓迎なので仲良くしてください。 |
| 提案手法 ($ratio = 0.3$) | 私も呼びタメ大歓迎です。なんて呼べばいいですか？ |
| 提案手法 ($ratio = 0.4$) | 私も呼びタメ大歓迎です。なんて呼んだらいいですか？ |
| 入力文 | お出かけからただいまー。草刈りとアメリカシロヒトリの駆除を頑張りました毛虫めー。 |
| 正解文 | ＜人名＞さん、お帰りなさい。お疲れ様でした。 |
| ベースライン | お大事にしてください。 |
| 提案手法 ($ratio = 0.1$) | お疲れ様でした。今日はゆっくり休んでくださいね。 |
| 提案手法 ($ratio = 0.2$) | おはようございます。今日も暑かったですね。 |
| 提案手法 ($ratio = 0.3$) | おはようございます。今日もお疲れ様 でした。 |
| 提案手法 ($ratio = 0.4$) | おはようございます。今日はお疲れ様です。 |
| 入力文 | 家庭科でやったけど髪の色と目の色と肌の色が関係するんやって服って。 |
| 正解文 | ファッション難しいのね。 |
| ベースライン | そうなんですね。お大事にしてくださいね。 |
| 提案手法 ($ratio = 0.1$) | 今髪染めてるんですか？ |
| 提案手法 ($ratio = 0.2$) | そーなんですか!?染めてないのか...。染めてないのか...。 |
| 提案手法 ($ratio = 0.3$) | そうなんですね！ありがとうございます。派手な色が好きなんですよね。 |
| 提案手法 ($ratio = 0.4$) | そーなんですね！そうなんですね！私も嬉しいです。 |
| 入力文 | ヒマだぁぁぁー。いっぱいパンケーキ焼いたー！ |
| 正解文 | パンケーキ、すげー！ |
| ベースライン | おめでとうございますー！ |
| 提案手法 ($ratio = 0.1$) | 美味しそうだよね。ザンビ食べたい。 |
| 提案手法 ($ratio = 0.2$) | 美味しそうだよね笑。 |
| 提案手法 ($ratio = 0.3$) | 一緒に食べようぜ。 |
| 提案手法 ($ratio = 0.4$) | ムキムキになったんだね。 |