外部知識を用いた雑談対話システムの汎化性能向上の検討 *

☆麻生大聖, 高島遼一, 滝口哲也, 有木康雄 (神戸大)

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，ユーザ発話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。入力発話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，発話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。

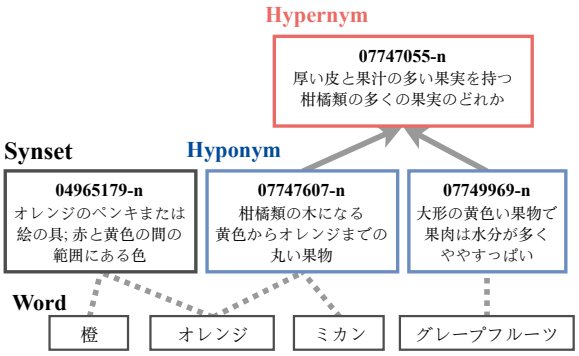

Fig. 1 Japanese WordNet

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

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，二発話のペアに分割した。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，名詞を含む発話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。

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[3] 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。

---

*Improvement of Generalization Performance of Non-task-oriented Dialogue System by Use of External Knowledge. by ASO, Taisei, TAKASHIMA, Ryoichi, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

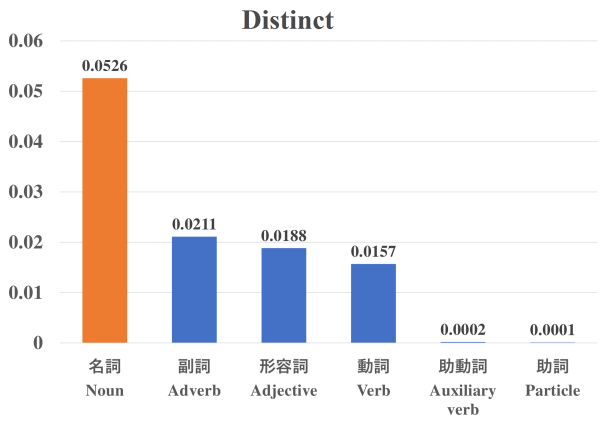

Fig. 2　Distinct of each part of speech in Twitter dialogue corpus

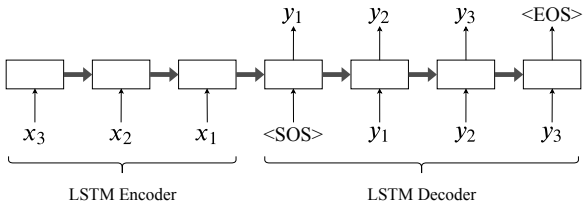

Fig. 3　LSTM Encoder-Decoder

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[4] 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。この手法をベースラインとする。

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，(1) 式から (6) 式により $V(w)$ に変換して入力する手法を提案する。(4)(5)(6) 式では，日本語 WordNet を検索している。

この手法では，名詞の上位概念に含まれる単語の Word2Vec による分散表現を，上位になるほど重みを小さくしながら，繰り返し加算していく ((2) 式)。上位概念が複数存在したり，概念に複数の単語が含まれる場合は，それらの分散表現の平均ベクトルをとる。提案手法による単語の分散表現の導出の概略図を Fig. 4 に示す。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

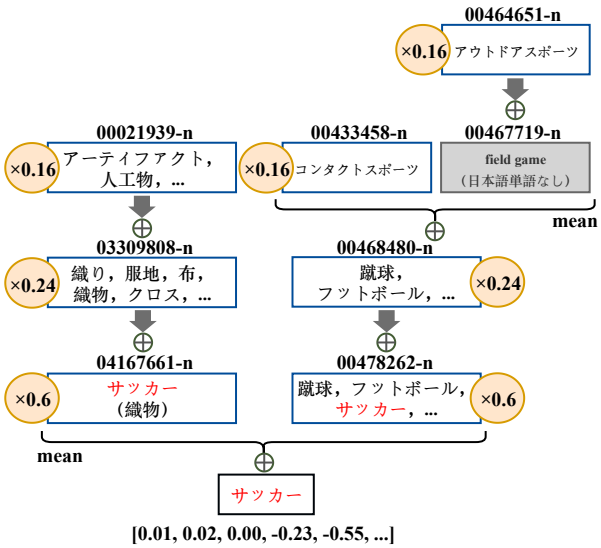

Fig. 4　Proposed Method ($ratio = 0.4, depth = 2$)

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。しかし，日本語 WordNet には収録されていない名詞が入力されることがあり，その場合には名詞以外の品詞と同様に Word2Vec による分散表現に変換して入力した ((1) 式)。

この手法には二つのパラメータ $ratio$ と $depth$ がある。$ratio$ は上位概念を加算する割合を示し，$ratio = 0$ 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し，$depth = 0$ のときには $ratio = 0$ のときと同じである。

5　実験

5.1　実験条件

Word2Vec による単語の分散表現の次元数は 256 とし，Skip-gram モデルで 10 回学習した。出現回数が 5 回未満の単語は除外し，Word2Vec の語彙数は 250,908 になった。

LSTM Encoder-Decoder のユニット数は 256，隠れ層は 3 層とした。学習用対話データの全ての目標文中に 5 回以上現れる単語のみを Decoder の出力単語の候補とすることで，出力語彙数が 32,302 となった。最適化手法には Adam を用いて初期学習率は 1e-4 とした。バッチサイズを 256 として 300 エポック学習した。過学習を抑制するために，ドロップアウト率を 20%とした。評価時には幅 15 のビームサーチ探索によって応答文を生成した。

提案手法において，上位概念を加算する割合を $ratio = 0.4$，加算する上位概念の最大の深さを $depth = 2$ と設定した。

$$V(w) = \begin{cases} (\text{単語 } w \text{ の } Word2Vec \text{ による分散表現}) & (\mid W2S(w) \mid = 0) \\ \frac{1}{|W2S(w)|} \cdot \sum_{s \in W2S(w)} SV(s, depth),\ depth \geq 0 & (otherwise) \end{cases} \tag{1}$$

$$SV(s,d) = \begin{cases} WV(S2W(s)) & (\mid S2H(s) \mid = 0\ or\ d = 0) \\ \frac{1}{|S2H(s)|} \sum_{h \in S2H(s)} SV(h,d) & (\mid S2W(s) \mid = 0) \\ (1 - ratio) \cdot WV(S2W(s)) & \\ \quad + ratio \cdot \frac{1}{|S2H(s)|} \cdot \sum_{h \in S2H(s)} SV(h, d-1),\ ratio \in [0,1] & (otherwise) \end{cases} \tag{2}$$

$$WV(ws) = \frac{1}{\mid ws \mid} \cdot \sum_{w \in ws} (\text{単語 } w \text{ の } Word2Vec \text{ による分散表現}) \tag{3}$$

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

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

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

5.2　分散表現の分布の比較

Word2Vec による分散表現と提案手法による分散表現に対して，主成分分析を行った。結果を Fig. 5 と Fig. 6 に示す。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。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。

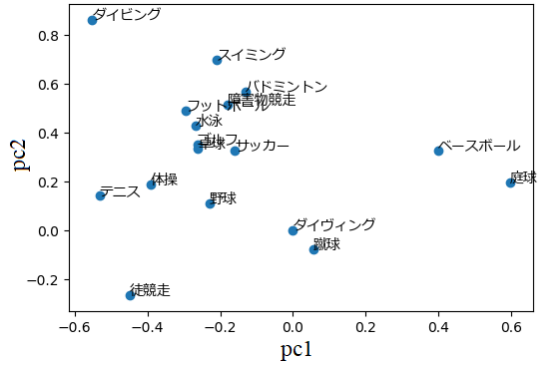

Fig. 5　PCA of Word2Vec distributed representation

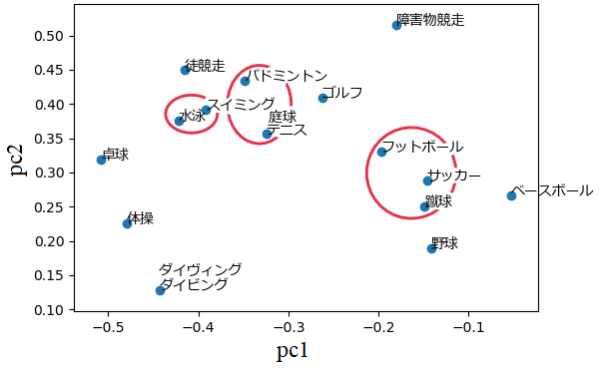

Fig. 6　PCA of proposed distributed representation

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 1 で各手法による応答文の BLEU[5] を比較している。提案手法では，ベースライン手法よりも BLEU が向上した。しかし，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を検討している。

Table 1　BLEU of each method

	ベースライン手法	提案手法
BLEU-1	0.128396	0.129586 (+0.9%)
BLEU-2	0.019919	0.02105 (+5.7%)

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 2 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，ベースライン手法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 (Fig. 5)。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 (Fig. 6)。例えば，

Table 2　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？
ベースライン手法	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	私も呼びタメ大歓迎なので気軽に絡んでください。なんて呼んだらいいですか？
入力文	オリーブ苦手なの？
正解文	苦手ってかアレってカラスの餌でしょ？
ベースライン手法	そーなんだ。
提案手法	果物全般の方が好き。
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。
ベースライン手法	お大事にしてください。
提案手法	お疲れ様でしたー。
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。
ベースライン手法	そうなんですね。お大事にしてくださいね。
提案手法	あ、そうなんですね。私も染めたことないんですけど、メイクしてないんですか？

「オリーブ」は比較的出現回数の少ない単語であったが，上位語に「核果」や「果実」のような単語をもつため，共通の上位語をもつ「さくらんぼ」や「桃」などの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，発話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。ベースライン手法よりも 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[6] 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。

参考文献

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

[4] Ilya Sutskever *et al.*, "Sequence to Sequence Learning with Neural Networks," In *Advances in neural information processing systems*, pp. 3104-3112, 2014.

[5] 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.

[6] Sascha Rothe and Hinrich Schutze, "AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes," *Proc. of ACL 2015*, pp. 1793-1803.