修士学位論文要約（令和４年３月）

加法構成性を持つ文ベクトルを用いた用例ベース対話システムの拡張に関する研究

矢作　凌大
指導教員：伊藤　彰則

Research on the Example-based Dialogue System Using Sentence Vectors with Additive Compositionality

Ryota YAHAGI
Supervisor: Akinori ITO

An example-based dialogue system is sometimes used, in which the dialogue is based on a set of example-response pairs, where the input is assumed by the developer and the response is set. While neural-based response generation can generate flexible responses due to the generalization capability of neural nets, it is difficult for the developer to design what kind of response should be made to the input. we proposed a response generation method using the additive compositionality of sentence vectors to generate sentences based on examples as an extension of the example-based dialogue system, and a response generation method using machine learning to ensure sentence generation based on a response of an example. This study suggests the possibility of realizing a dialogue system that generates responses based on examples.

1. はじめに

近年、あらゆる入力発話に返答できるニューラルネットワークを用いた対話システムの研究が盛んである。ニューラルベースの対話システムは応答が設計できないため、製品化やサービス化する際には用例ベースの対話システムのように開発者が応答を設計できるものが望ましい。しかし用例ベースで幅広い話題に対応するには開発者の負担が非常に大きい。そこで本研究では、文を固定次元のベクトルにエンコード、またベクトルから文をデコードするモデルの構築を行い、それを用いてニューラルネットワークによって獲得された単語ベクトルに見られる、ベクトルの加算や減算により対応した意味の計算が為される性質の加法構成性 [1]を文単位のベクトルに応用することで、想定外の入力発話にも用例を基にした応答を行う対話システムの検討を行う。加えて類似した応答対である応答対Aと応答対Bをペアとして収集し、A の入力発話ベクトルと B の入力発話ベクトルの差分とBの応答発話を利用して応答Aを生成する学習を提案モデルに行わせる。

2. 文の固定次元ベクトルへのエンコード

文を固定次元のベクトルにエンコードすることができれば，文ベクトル同士の加減算が可能となるため，用例応答対における入力発話と実際の入力発話の差分ベクトルを用例応答対における応答のベクトルに加算する図 1 に示すような応答生成手法が考えられる。

この手法には文を固定次元のベクトルにエンコード、またベクトルから文をデコードするモデルが必要となる。そこで近年の自然言語処理にて利用されるTransformer [2]のエンコーダ出力が可変長 n であるのに対し Query が一定数 m の Attention を導入することで入力を固定次元にエンコードするオートエンコーダである m-Query を構築した。同モデルは通常のオートエンコーダタスクのように入力をそのまま復元するタスクは行わず、より複雑な入力単語の一部をマスクするマスク付き文復元を学習するにも関わらず、エンコーダ出力を平均する従来のオートエンコーダよりも高精度に文を復元可能である。

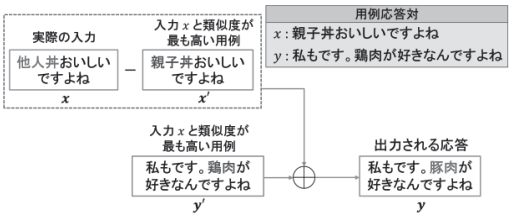

図 1　加法構成性を利用した応答生成手法

3. 差分ベクトルにゲインを与えることの有効性

図 1 に示す提案手法では殆ど期待通りの文生成は確認されなかった。この原因は類似した文における差分がそのほかの成分に埋もれてしまうことで、文単位の差分ベクトルの絶対値が小さくなりすぎてしまうことが挙げられる。そこで、差分ベクトルにゲインを

与えることで差分ベクトルを増幅させ、用例応答対の応答ベクトルに加算したベクトルをデコーダに入力することで文を生成する。ゲインを変化させながら期待した出力との一致度である BLEU を算出し、結果は図 3 の通りとなった。このとき学習した文の数は約 26 億文である。

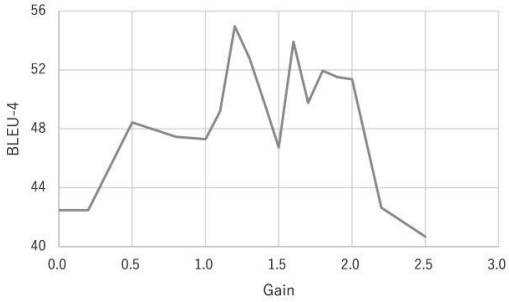

図 3　ゲインと BLEU の関係

同図よりゲインを与えない 1.0 の場合より顕著に BLEU スコアが上昇している点があることがわかる。BLEU が特に上昇したゲインにおける生成文の例を表 1 に示す。同表より入力の差分ベクトルに適切なゲインを与えることで、想定した出力を得られる場合もあるが、ゲインによるばらつき非常に大きいことが問題点として挙げられる。

表 1　ゲインによる生成文の変化

input (expected)	gain	generated
A woman is standing.	1.0	of woman does that for a while.
− A man is standing.	1.2	and woman does that for a while.
+ The man does that for a while.	1.6	and woman does that for a while.
(The woman does that for a while.)	2.5	**the woman does that for a while.**
A woman was standing.	1.0	they woman does that for a while.
− A man is standing.	1.2	**of woman did that for a while.**
+ The man does that for a while.	1.6	of woman then that for a while.
(The woman did that for a while.)	2.5	it woman newly that for a while.
Were you hungry?	1.0	went am notangry.
− Are you angry?	1.2	were am not angry.
+ I am not angry.	1.6	**i was not hungry.**
(I was not hungry.)	2.5	knew was night dessert.

* ボールド体は期待した出力とほぼ同様の生成文.

4. 適切な応答生成を保証するための学習

前章まで適切な応答文が得られづらい最大の原因は、モデルの学習目的と期待する生成文に大きな乖離が存在することである。そこで本章では構築したオートエンコーダを事前学習モデルとして、同モデルに類似した応答対である応答対 A と応答対 B のペアを収集した Example Pairs on Semantic Similarity (EPoSS)データセットを用いて、応答対 A の入力発話から応答対 B の入力発話の差分ベクトルと、応答対 B の応答のベクトルを利用して応答対 A の応答を生成する学習を行わせる。

EPoSS データセットの構築には大規模対話コーパスである LCCC [3]を利用する。LCCC は中国語であるため、機械翻訳を用いて英訳した。英訳されたコーパスに対して Sentence BERT [4]により入力発話とその応答をベクトル化し、入力発話同士と応答の発話同士どちらにおいてもコサイン類似度が 0.7 以上となる応答対を収集することで EPoSS データセットを構築した。このとき条件を満たす対話をすべて収集したデータを All、All では挨拶のような定型文同士の対話が非常に多く抽出されてしまいデータが偏るという懸念から、類似対話を複数持つ対話に対してはその内ランダムにひとつの対話を抽出したデータを Balanced として 2 種のデータセットを構築した。

学習率(LR)や学習に用いるデータセット、モデル構造を変化させて EPoSS の評価セットに対して BLEU スコアを算出した結果を表 2 に示す。Example は基となる対話における応答をそのまま出力する手法、Vanilla は単純な加減算を行う手法、Calc は Vanilla の手法をそのまま EPoSS によって学習させたモデル、Cat は入力における差分ベクトルと基となる応答のベクトルをデコーダに入力して EPoSS で学習したモデルである。同表から目的を保証した学習が必要であり、設計した用例を基に応答を生成することが可能であることが示唆される。

表 2　各手法による BLEU スコア

Model	Corpus	LR	BLEU incl. Err		BLEU exc. Err	
			ex	sim	ex	sim
Example	-	-	100.0	32.14	100.0	32.14
Vanilla	-	-	26.66	11.29	30.15	12.77
Calc	All	4e-4	6.87	7.74	11.95	13.47
		4e-5	11.10	10.49	18.58	17.55
	Balanced	4e-4	23.62	23.40	24.29	24.07
		4e-5	22.95	20.31	27.15	24.03
Cat	All	4e-4	16.60	16.43	20.89	20.67
		4e-4	12.14	12.01	22.25	22.03
	Balanced	4e-4	27.36	23.96	28.36	24.83
		4e-5	**28.99**	**26.03**	**33.53**	**30.10**

5. まとめ

文ベクトルの加法構成性を利用した応答生成手法を提案しある程度の有効性は確認されたものの、ゲインによるばらつきが非常に大きいことが懸念された。そこで適切な応答生成を保証した学習を行うことで、設計した応答を基に新しく応答を生成可能であることを明らかにした。

文献

1) T. Mikolov et al., ``Distributed representations of words and phrases and their compositionality." In Proc. of NIPS, pp. 3111-3119, 2013.
2) A. Vaswani et al., ``Attention is all you need.'' In Proc. NIPS, pp. 5998-6008, 2017.
3) Y. Wang, et al., ``A large-scale chinese short-text conversation dataset.'' In Proc. of NLPCC, pp. 91-103, 2020.
4) N. Reimers et al., ``Sentence-bert: Sentence embeddings using siamese bert-networks.'' In Proc. of EMNLP-IJCNLP, pp. 3982-3992, 2019.