与えた発話意図の再予測を用いた応答生成モデルの検討

隆辻 秀和[1*]　吉野 幸一郎[1]　須藤 克仁[1]　中村 哲[1]
Takatsuji Hidekazu[1]　Yoshino Koichiro[1]　Sudoh Katsuhito[1]　Nakamura Satoshi[1]

[1] 奈良先端科学技術大学院大学 先端科学技術研究科
[1] Graduate School of Science and Technology, NARA Institute of Science and Technology

Abstract: 言語生成は，与えられた外部情報のセットに対して，自然言語文をドメインに適当な形で生成するタスクである．近年，言語生成に用いられるニューラルネットワークを用いた手法は，より自然で柔軟な応答生成が実現できることが知られている．一方で，入力となる外部情報に対応する文生成を単語予測のモデルで行うため，モデルがどの情報を利用し文を生成したか，生成時に与えた情報が反映されているか知ることが難しい．そこで本研究では，与えた外部情報を生成文に反映することを保証するため，与えた外部情報を再予測するモデルと再予測の結果に対する損失を利用した．アノテーション済みのコーパスを用いた実験を行い，生成された文に対して評価を行った．

1　はじめに

言語生成は自然言語処理における重要なタスクの一つである．言語生成器は生成で考慮すべき外部情報を入力として受け取り，ドメインや文脈に合わせて適切な文生成を行う．機械翻訳であれば，翻訳元の言語で記述された文が与えられ，意味を損なわないように翻訳先の言語で記述した文を生成する．また，質問応答システムではユーザーから投げかけられた質問文の他に，システムが保持している知識などの追加情報も用いて回答を生成する．

言語生成の手法として，テンプレートを用いる手法やルールに基づく手法などが考えられてきた [1, 2, 3] が，近年ではニューラルネットワークを用いた手法が一般的になっている．ニューラルネットワークを用いた手法は，従来の手法に比べて流暢な文生成を可能であることが，さまざまな研究によって知られている．最もよく知られているモデルとして Sequence-to-Sequence (Seq2Seq)[4] がある．このモデルでは言語生成を入力の単語系列から出力の単語系列への写像関係を学習する問題として捉えることで，学習時に存在しない未知の系列に対しても従来手法に比べて自然な生成を可能にしている．

入力文以外の外部情報を考慮するような言語生成モデルはさまざまに提案されている．Li ら [5] は出力文に反映すべき言語的特徴として個人性を取り扱う手法を提案している．この手法では，ある個人を表現する分散表現をデコーダの各ステップで直前の単語と共に入力することで，条件付き生成を実現している．

*連絡先：奈良先端科学技術大学院大学 先端科学技術研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: takatsuji.hidekazu.sx1@is.naist.jp

また，Eric ら [6] は Seq2Seq モデルのエンコーダに，質問文と slot-value 形式で与えられる外部情報の slot の情報を入力し，コピーメカニズムを利用して外部情報を出力系列に反映する手法を提案している．

これらのモデルは，生成する単語系列に対する予測誤りを元に学習を行うため，追加で与えた情報が適切に考慮されていることが保証されないという問題がある．この問題は，特に，クエリの他に与えられる外部情報が出力に含まれることが期待される状況では大きな課題となる．

本研究では，モデルに入力文の他に与えた外部情報が生成文に含まれることを保証するための条件付き言語生成モデルを提案する．具体的には単語系列の予測に加えて，与えた外部情報の再予測を行うモデルを導入し，外部情報の予測結果に対して損失を計算する．これによってモデルの単語生成時に外部情報が必ず参照する状態に含まれていることを保証し，生成される単語列に対しても外部情報が適切に含まれることを期待した．提案するモデルの有効性を確認するために，先行研究 [6] と同じコーパスを用いて実験を行い，結果を確認した．

2　関連研究

ニューラルネットワークを用いた言語生成において，入力文以外の情報を考慮して生成を行う研究は多く存在する．Seq2Seq によるモデルの他に，メモリネットワークと呼ばれる Key-Value 構造を行列で表現したネットワークによって外部情報を考慮するモデルが広く知られている．Sukhbaatar ら [7] は外部情報を Key-Value からなる辞書オブジェクトとみなし，入力の質問文に

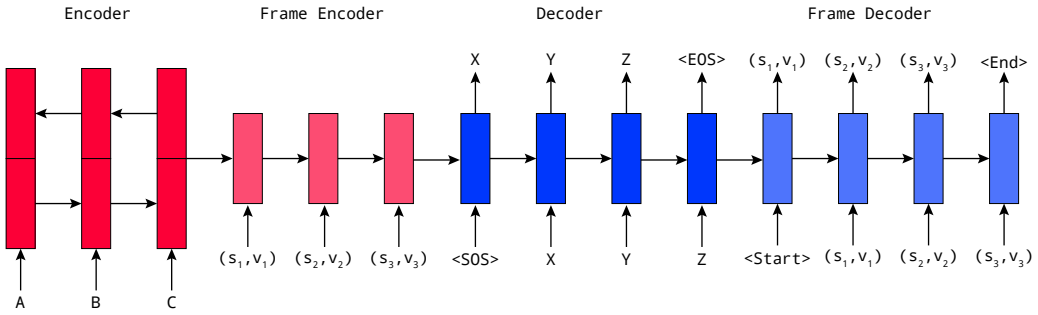

図 1: 提案モデルの概要

対して内積注意を計算することで，外部情報を参照した応答生成を行うモデルを提案している．このモデルは，回答が一つの単語に定まるような質問応答タスクにおいて有効性が確認されている．また，Madotto ら [8] は Sukhbaatar らのモデルの拡張を行い，対話履歴をメモリに入力し，出力として単語の系列を予測するモデルを提案している．

値を明示的に埋め込む手法としては，Qian ら [9] による Seq2Seq を拡張し生成文中に与えられた外部情報の値を埋め込む手法が知られている．この研究では，外部情報を生成に利用するか決定する予測器を学習し，予測器の出力に応じて利用する生成器を切り替える．この手法では与えられた外部情報の特定の Key-Value の組しか考慮できないという問題があるが，一方で指定した値が必ず出力されるように生成を実行しているため，出力に Key-Value の値を反映することを従来モデルに比べてより良い形で保証することができる．

これらの先行研究に対して提案モデルでは，目的の異なる複数のモデルを同時に学習させることで，生成タスクの精度を向上させることを意図している．このような学習はマルチタスク学習と呼ばれ，系列モデルに対してマルチタスク学習を適用する手法は既に Luong ら [10] によって研究されている．Luong らの研究ではエンコーダとデコーダに対して 1 対多や多対多のモデルについて取り上げており，これらのケースでは個々のエンコーダあるいはデコーダについては並列の関係となっている．一方で，本研究では複数のエンコーダ及びデコーダを直列に取り扱っており，この点において先行研究と異なる．

3 条件付き応答生成モデル

図 1 に提案するニューラル文生成モデルについての概要を示す．提案モデルは Seq2Seq の枠組みを条件付き生成に拡張したものであり，外部情報を与えるデコーダ及び与えた外部情報を生成後に再予測するネットワークを追加している．

本研究では条件付きの応答生成タスクについて取り扱う．このタスクでは，クエリ $\mathbf{q} = (q_1, q_2, \ldots, q_N)$，応答文 $\mathbf{r} = (r_1, r_2, \ldots, r_M)$ の組の他に，応答文に含まれるべき外部情報の集合 $\mathcal{F} = \{(s_1, v_1), (s_2, v_2), \ldots, (s_K, v_K)\}$ が与えられる．応答生成モデルは確率分布 $P(\mathbf{r} \mid \mathbf{q}, \mathcal{F})$ を訓練データに対して尤度最大となるように学習を行う．

提案モデルでは上述の確率分布を尤度最大となるように学習する他に，生成後の隠れ状態を利用して与えた外部情報の再予測を行う．これは，生成時の各ステップの隠れ状態が与えた外部情報を保持していることが保証を行うことで，生成結果に対して与えた外部情報の値が反映される可能性が高くすることができるのではないかと考えたためである．

3.1 エンコーダ

Seq2Seq では与えられるクエリを Recurrent Neural Network (RNN) を用いて分散表現に変換する．クエリの各トークン q_t に対応する隠れ状態 h_t は次式によって得られる．

$$\overrightarrow{h}_t^{enc} = \mathrm{RNN}(q_t, \overrightarrow{h}_{t-1}^{enc}) \tag{1}$$

$$\overleftarrow{h}_t^{enc} = \mathrm{RNN}(q_t, \overleftarrow{h}_{t+1}^{enc}) \tag{2}$$

$$h_t^{enc} = \mathrm{concat}(\overrightarrow{h}_t^{enc}, \overleftarrow{h}_t^{enc}) \tag{3}$$

本研究では RNN のユニットとして LSTM を利用した．また，クエリをエンコードする際には，双方向 RNN を用いた．

3.2 フレームエンコーダ

今回取り扱うタスクではクエリ $\mathbf{q}$ の他に外部情報 $\mathcal{F}$ の情報を入力する必要がある．提案モデルでは，外部情報もまた RNN を用いて分散表現へと変換し，利用する．外部情報は集合となっているため，RNN に入力するために何らかの順序を用意する必要がある．提案

モデルでは，外部情報の集合に含まれる各要素に対して擬似的に順序を付与した．

$$h_k^{fenc} = \mathrm{RNN}((s_k, v_k), h_{k-1}^{fenc}) \tag{4}$$

3.3 デコーダ

デコーダでは，与えられたクエリと外部情報から応答となる単語列 $\mathbf{r}'$ を生成する．生成の各ステップでは，直前の予測単語 r'_{t-1} と隠れ状態 h_{t-1}^{dec} から次の単語 r'_t を生成する．学習時には直前の単語として予測単語 r'_{t-1} でなく正解系列の単語 r_{t-1} を用いる．

$$h_t^{dec} = \mathrm{RNN}(r'_{t-1}, h_{t-1}^{dec}) \tag{5}$$

$$r'_t = \mathrm{softmax}(\boldsymbol{W}_o^{dec} h_t^{dec}) \tag{6}$$

3.4 フレームデコーダ

デコーダが単語列 $\mathbf{r}'$ を生成する際に，その隠れ状態に外部情報 $\mathcal{F}$ が含まれることを保証することで生成される単語列の品質を改善することが本研究の目的である．この目的のために，提案モデルでは単語生成後に与えた外部情報について予測を行うネットワークを追加する．このネットワークもまた RNN を用いるため，外部情報 $\mathcal{F}$ を順番に予測しなければならない．提案モデルでは，フレームエンコーダに外部情報が与えられた順序と同じ順序で外部情報を予測するモデルとした．

$$h_t^{fdec} = \mathrm{RNN}((s'_{t-1}, v'_{t-1}), h_{t-1}^{fdec}) \tag{7}$$

$$(s'_t, v'_t) = \mathrm{softmax}(\boldsymbol{W}_o^{fdec} h_t^{fdec}) \tag{8}$$

3.5 損失関数

提案モデルは，単語の予測誤りに基づく損失と外部情報の予測誤りに基づく損失の二つを学習に用いる．単語予測についての交差エントロピー損失 $\mathcal{L}_w$ および外部情報予測についての交差エントロピー損失 $\mathcal{L}_{fr}$ の重みつき線形和 $\mathcal{L} = \mathcal{L}_w + \alpha\mathcal{L}_{fr}$ を損失関数として利用する．

4 評価実験

提案モデルとベースラインモデルを用いて，応答生成実験を実施し，自動評価尺度と主観評価を用いて提案モデルが生成する応答の品質が改善されたか評価した．実験において比較対象となるベースラインモデルは，提案モデルからフレームデコーダを省いたモデルとした．これは損失関数の重みパラメータ α を 0.0 にすることと等しい．

4.1 実験設定

実験では DSTC2[11] において配布されたコーパスを利用した．このコーパスはレストラン情報案内を行うシステムと利用者の対話を収録しており，システムと利用者どちらの発話に対しても対話行為と発話に含まれる情報が slot-value 形式で付与されている．DSTC2 におけるシステムはテンプレートを利用して発話が生成されているため，実験ではシステムの発話をクエリ，利用者の発話に付与された slot-value の組を外部情報として利用者の発話を生成するタスクを行った．slot-value 形式で提供される情報は 255 種類あり，それぞれがレストランの値段や提供する料理など，レストランを選ぶ時に用いる情報を表現している．コーパスには合わせて 16,551 件の発話ペアが含まれており，このコーパスを 15,723/414/414 に分割し，それぞれを訓練用，開発用，テストデータとして利用した．

学習にあたって，DSTC2 コーパスは含まれる発話ペアの数が少ないため，応答生成の学習に不十分である可能性を考え，Reddit[1]より収集した観光情報に関する 50 万件の発話ペアを事前学習に利用した．

それぞれのコーパスに対する前処理として SentencePiece[12] を用いたトークナイズを行った．トークナイザの学習には Reddit より収集したコーパスを利用し，単語サイズは 16,000 とした．

学習時には最適化手法は Adam[13] を利用した．また，損失関数の重みパラメータ α は 1.0 に固定して学習を行った．

4.2 参照応答を用いた自動評価

評価尺度として，BLEU[14]，パープレキシティ，Entity. F1, および提案モデルのみを対象として外部情報の再予測精度を用いた．

BLEU は主に機械翻訳における翻訳の評価を行うために用いられるが，複数の先行研究 [5, 6] においてモデルが生成した応答を評価する手法として用いられている．生成結果が正解系列に近ければ近いほど高い値となる．本研究では，モデルが生成した文ごとに BLEU スコアの計算を行い，その平均値をモデルの BLEU として利用した．

Entity. F1 は先行研究 [6] においてモデルが与えた外部情報を考慮しているか確認するために用いられた指標である．生成された結果が与えられた外部情報を含

[1] https://reddit.com

表 1: 自動評価尺度による結果

Model	BLEU	Ppl.	Ent.F1	Accuracy
baseline	55.42	2.1890	69.15	-
proposed	50.85	2.0127	69.82	0.8561

表 2: 主観評価における平均スコア

Model	Score (avg.)
baseline	2.2894
proposed	2.2206 (-0.0678)

表 3: 生成例の比較

query	What kind of food would you like?
response	im looking for a restaurant in the north part of town serving kosher food
frame	{(food, kosher), (area, north)}
baseline	north part of town
proposed	i want a restaurant in the north part of town serving kosher food
query	Hello, welcome to the Cambridge restaurant system? You can ask for restaurants by area, price range or food type. How may I help you?
response	i am looking for a restaurant in the south part of town and it should serve cantonese food
frame	{(food, cantonese),(area, south)}
baseline	im looking for a restaurant in the south part of town that serves
proposed	im looking for a restaurant in the south part of town serving *romanian* food

んでいるかについて，マルチクラスに拡張した F1 スコアを用いて計算する．モデルが生成結果に外部情報を正しく含んでいれば，より高い値となる．

4.3 主観評価実験

生成される応答が指定された外部情報を含むものとなっているかどうかを判断するために Amazon Mechanical Turk[2]を利用して主観評価実験を行った．主観評価実験ではベースラインモデル，提案モデルいずれのモデルから生成された応答であるかを伏せた上で，評価者にクエリ，外部情報および生成された応答の三つ組を提示し，生成された応答が外部情報として与えられる値を正しく反映しているかについて三段階で評価するように指示した．各生成結果に対して 3 名の評価者による評価を実施し，3 名のスコアの中で過半数を占めるスコアをその生成結果に対するスコアとして採用した．過半数となるスコアが存在しなかった場合はスコアを 2 として処理を行った．

5 実験結果

表 1 に自動評価尺度における評価結果を，表 2 に主観評価における各モデルの平均スコアを示す．自動評価尺度を用いた評価結果では，提案モデルがベースラインモデルに比べて大幅に BLEU を下げていることが見て取れる．その一方で，パープレキシティはベースラインモデルに比べてやや低い値となっている．このことから，提案モデルはより良い言語モデルを学習しているが，その生成結果がベースラインと比べてより正解となる単語列に近いものとはなっていないことが示されている．

また，Entity. F1 の改善幅は非常に小さく，ベースラインと提案モデル間ではほぼ差がない．その一方で，外部情報の再予測精度は 85.6 % とチャンスレートを大幅に上回っており，外部情報がデコーダの隠れ状態には含まれていることを示唆している．このことから，提案モデルは意図通りに動作しているが，必ずしも生成品質の改善に寄与しているわけではないということがわかる．

次に，表 2 に主観評価の結果を示す．表 2 の結果におけるモデル間のスコアの差は非常に小さい．このことからも提案モデルがベースラインに比べて何らかの改善を達成したということは認められなかった．

最後にそれぞれのモデルの生成例について比較する．表 3 にベースライン及び提案モデルの生成例をそれぞれ示す．表上部に示す生成例は，提案モデルがベースラインに対してより良い生成となっている例である．この生成例では，ベースラインは与えられた外部情報の中で area しか反映できていないが，提案モデルではどちらの情報も正しく反映することができている．一方で，下部に示す生成例ではどちらのモデルも正しく情報を反映することができていない．ベースラインの生成例は先ほどと同様に area に関する情報のみが反映できており，他の情報は反映することができていない．提案モデルの生成例はどちらの情報に関しても，指定された slot に関する情報を反映しているが，food に関し

[2] https://www.mturk.com

ては指定された cantonese ではなく誤った romanian を生成している．この例からも提案モデルが今回取り組んだタスクに対して，必ずしも良いモデルとなっているとは言えない．

6 おわりに

本研究では，与えられた外部情報を含むことを期待するような条件付き言語生成において，生成結果の改善を目的として新たなモデルを提案した．提案モデルではデコーダで生成を行った後に，言語生成モデルに与えた外部情報を予測することによって，生成の段階で与えた外部情報を隠れ状態が含むように促すことで生成品質の改善を期待した．提案モデルの有効性を確認するために，予測を用いないベースラインモデルと提案モデルを用いて生成品質について，複数の評価尺度を用いた評価と，クラウドソーシングを用いた主観評価を実施した．これらの結果から，提案モデルは意図した動作を行っているものの，生成結果の改善に繋がらないということが示された．考えられる要因の一つとして，言語表層における制約として外部情報を考慮していないため，モデルが表層を正しく生成するように学習しなかったということを考えている．これを改善する手法として，生成時に単語を直接出力するのではなく，slot-value 形式の slot を出力し外部情報で指定された値をコピーして利用するなどを検討する必要がある．

参考文献

[1] R. Dale, S. Geldof, and J.-P. Prost. “CORAL: Using Natural Language Generation for Navigational Assistance”. *Proceedings of the 26th Australasian Computer Science Conference*. Vol. 16, pp. 35–44. (2003).

[2] 山崎 健史 et al. “フローグラフからの手順書の生成”. **情報処理学会論文誌** 57, 3, pp. 849–862. (2016).

[3] R. Kondadadi, B. Howald, and F. Schilder. “A Statistical NLG Framework for Aggregated Planning and Realization”. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1406–1415. (2013).

[4] I. Sutskever, O. Vinyals, and Q. V. Le. “Sequence to Sequence Learning with Neural Networks”. *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al., pp. 3104–3112. (2014).

[5] J. Li et al. “A Persona-Based Neural Conversation Model”. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 994–1003. (2016).

[6] M. Eric and C. Manning. “A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue”. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 2, pp. 468–473. (2017).

[7] S. Sukhbaatar et al. “End-To-End Memory Networks”. *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al., pp. 2440–2448. (2015).

[8] A. Madotto, C.-S. Wu, and P. Fung. “Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems”. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1468–1478. (2018).

[9] Q. Qian et al. “Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation”. en. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 4279–4285. (2018).

[10] M.-T. Luong et al. “Multi-task Sequence to Sequence Learning”. *the 4th International Conference on Learning Representations (ICLR)*. (2016).

[11] M. Henderson, B. Thomson, and J. D. Williams. “The Second Dialog State Tracking Challenge”. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 263–272. (2014).

[12] T. Kudo and J. Richardson. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71. (2018).

[13] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. *the 3rd International Conference on Learning Representations (ICLR)*. (2015).

[14] K. Papineni et al. “BLEU: A Method for Automatic Evaluation of Machine Translation”. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318. (2002).