生成と分類のマルチタスク学習による感情が考慮された対話応答生成

井手竜也
早稲田大学基幹理工学部
t-ide@toki.waseda.jp

河原大輔
早稲田大学基幹理工学部
dkw@waseda.jp

1 はじめに

計算資源などの拡大に伴う、自然言語処理の発展は著しいものである．翻訳や要約をはじめとするタスクで、今やコンピュータが人間に劣らない性能を示すことも珍しくない．同じような技術のもとで対話システムも進歩し、実用化が期待されつつある．

人間がコンピュータと自然に対話するためには、コンピュータに人間らしさが必要である．こうした対話を実現するための方法 [1] はいくつか提案されている．たとえば知識や常識にもとづいて対話を行うこと [2] や、自身および相手の個性に触れながら対話を行うこと [3] である．中でも本研究は感情という要素 [4] に焦点を当てる．

本研究は話者の感情が考慮された対話をマルチタスク学習によって実現する．また感情の階層的な関係 [5] に目を向け、粒度の異なるいくつかの感情認識を同時に学習させる．なおこのマルチタスク学習は類似タスクの間で相補的な情報共有を目指すものではなく、あえて感情認識の精度向上は考慮しない．またマルチタスク学習における感情認識の比率が過剰であるという懸念のもと、各ロスに重みづけを施すことでさらなる品質の改善を求める．事前学習済みの Transformer [6] モデルである BART [7] をもとにモデルを構築し、応答生成と感情認識のマルチタスク学習を行う．これは英語を対象に、文脈を伴わない対話で行う．

自動評価と人手評価から提案手法による応答生成の有効性を確認した．つまり応答生成と感情認識のマルチタスク学習が、生成される応答をより発話の感情が考慮されたものにすることがわかった．このことは感情という側面にかぎらず、流暢さや有益さといった品質も同時に改善された．またロスの重みづけによるパラメータの制御が、モデルの性能をより向上させることがわかった．

2 関連研究

感情にもとづく応答生成に関しては、ECM [8] が代表的である．ECM は Emotion Classifier とともに用いられ、与えられた感情に依存した応答を生成することができる．EmpTransfo [9] は本研究と類似したモデルである．GPT [10] による応答生成に Emotion と Action を適用させ、生成される応答を高品質なものにしている．

これらのモデルは応答がもつ感情に注目したもので、発話の感情をふまえて応答を生成するわけではない．一方で TG-EACM [11] は、発話に込められた感情と応答が抱くべき感情の双方を考慮するようなモデルである．モデルは与えられた発話から発話と応答の感情をともに推測するような分布を学習する．

ECM や TG-EACM は感情を理解するためのユニットを独立にもつが、本研究はそれを単一のモデルで完結させる．それによってパラメータの冗長性が削減され、より効率的な感情の理解および応答の生成が実現されることを期待する．

3 感情が考慮された応答生成モデル

3.1 概要

生成タスクとして応答生成を行い、分類タスクとして感情認識を行う．応答生成と感情認識をマルチタスク学習によって同時に学習することで、与えられた発話の感情を理解した上で応答を生成できるようになることを狙う．

マルチタスク学習はいくつかの類似したタスクを対象とすることが多い．これはタスク同士が互いに情報を共有し、それぞれのタスクが性能を向上すること期待するためである．しかし今回のマルチタスク学習はあくまでも応答生成の品質改善が目的であ

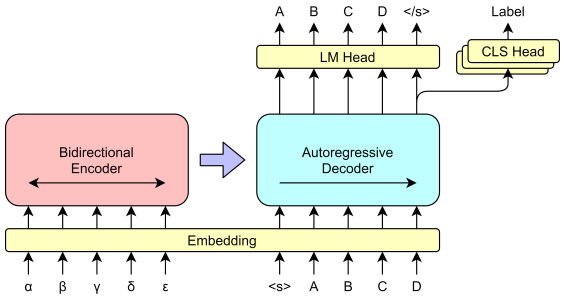

図 1　応答生成モデルのアーキテクチャ

り、感情認識の性能を向上することは考えない．これは一般的なマルチタスク学習と異なる点である．

モデルは BART [7] をもとに作成する．アーキテクチャは図 1 のとおりで、モデルは学習されるタスクごとに Head として出力層をもつ．これは応答生成で単語を生成するための LM Head と、分類タスクを解くための線形層である CLS Head からなる．なお CLS Head はたとえば Positive や Negative といったラベルを推測することで分類タスクを解く．CLS Head に関しては、分類タスクごとに 1 つを設ける．

各タスクの入出力形式は BART と同様である．生成タスクでは Encoder と Decoder にそれぞれ発話と右シフトされた応答を入れ、分類タスクでは Encoder と Decoder に発話と右シフトされた発話を入れる．

学習のアルゴリズムはおおむね MT-DNN [12] にしたがう．学習すべき各タスクはミニバッチごとに選択される．タスクによって異なるロスを計算し、ミニバッチごとにパラメータを更新していく．

3.2　タスクとロス

与えられる発話を $\boldsymbol{x} = (x_1, \ldots, x_M)$、モデルのパラメータを $\boldsymbol{\theta}$ とする．各タスクに対するロスをもとに $\boldsymbol{\theta}$ を更新することによって、モデルは学習される．

生成　$\boldsymbol{x}$ に対する応答を $\boldsymbol{y} = (y_1, \ldots, y_N)$ とすれば、$\boldsymbol{x}$ から適切な $\boldsymbol{y}$ を推測できるように学習する．

$$\mathcal{L}_{\text{gen}} = -\sum_{j=1}^{N} \log p(y_j | \boldsymbol{x}, y_1, \ldots, y_{j-1}; \boldsymbol{\theta}) \qquad (1)$$

分類　$\boldsymbol{x}$ の正解ラベルを c とすれば、モデルは $\boldsymbol{x}$ から c を推測する．

$$\mathcal{L}_{\text{cls}} = -\log p(c | \boldsymbol{x}; \boldsymbol{\theta}) \qquad (2)$$

3.3　ロスの重みづけ

提案のマルチタスク学習は生成と分類のタスクを同時に学習するものである．しかしそこには分類タスクに関する学習の比率が過剰である可能性がある．一般的な分類タスクを解くとき、学習の終了は Validation データによるロスの収束をもって定められることが多い．一方で本研究の対象は生成タスクであり、それに必要なエポック数は分類タスクのものよりも多い．

そこで各ロスに重みづけを施すことを考える．具体的には応答生成に対する重みを 1 に固定しつつ、感情認識の重みを 0 以上 1 以下で変動させる．それによりパラメータの更新に際する分類タスクの寄与を抑える．

表 1　各データセットのサンプル数

データセット	Train	Val.	Test
DailyDialog	76,052	7,069	6,740
Twitter Emotion Corpus	16,841	2,105	2,105
SST-2	16,837	872	1,822
CrowdFlower	15,670	1,958	1,958

4　実験

4.1　データセット

マルチタスク学習として、応答生成のほかに感情認識の 3 タスクを行う．それぞれは 6 ラベル・2 ラベル・12 ラベルによる分類タスクで、感情認識・CG 感情認識・FG 感情認識と呼ぶ．なお感情認識の各データセットは Unify Emotion Datasets [13] を参考に選択した．

応答生成　応答生成のデータセットとしては DailyDialog [14] を用いる．これはマルチターンによる対話コーパスであり、2 ターンずつを抽出することによって発話と応答の組を得る．これに収録されている各発話は感情ラベルを伴うが、本研究ではそれらを使わない．付与されている感情ラベルのほぼすべてが Other であり、このことは本研究の目的にそぐわないと判断したためである．

感情認識　軸となる感情認識のデータセットには Twitter Emotion Corpus [15] を用いる．Twitter のハッシュタグをもとに構築されたもので、{Anger, Disgust, Fear, Joy, Sadness, Surprise}の 6 ラベルからなる．なおこのデータセットには{Train, Validation, Test}の区別がない．よって全データセットの 80% を Train に割り当て、残りの 10%ずつを Validation と Test にする．

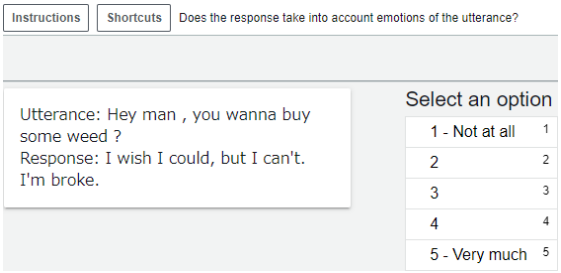

図 2　人手評価として行うクラウドソーシングの例

CG 感情認識　Coarse-Grained な感情認識については SST-2 [16] を用いる．これは映画のコメントに{Positive, Negative}のラベルが付与されたデータセットである．感情分類のサンプル数とバランスを保つべく、Train データのサンプル数を 25%に減らす．

FG 感情認識　Fine-Grained な感情分類には CrowdFlower が提供する感情ラベル付きコーパスを用いる．Empty というラベルを除外し、{Anger, Boredom, Enthusiasm, Fun, Happiness, Hate, Love, Neutral, Relief, Sadness, Surprise, Worry}の 12 ラベルによる分類タスクとして扱う．なおこれも Twitter Emotion Corpus と同じく{Train, Validation, Test}の区別をもたないため、全体を 8:1:1 に分割する．また SST-2 と同じ理由のもと、全体の 50%のみを抽出する．

各データセットにおけるサンプル数の内訳を表 1 にまとめる．

4.2　モデルの学習

学習に必要となるハイパパラメータは、BART の論文 [7] や Fairseq の例を参考に設定する．学習率は 3e-5 と定め、パラメータは Weight Decay を伴う Adam で最適化する．応答生成に関しては、ロスの Negative Log-Likelihood に 0.1 の Label Smoothing を施す．また入出力のトークン数は最大 64 に決め、学習は 64 エポック行うことにする．

生成時のハイパパラメータに関しても、上記を参考に設定する．5 ビームによるビームサーチで単語を選択し、ある n-gram が 3 個以上連続する場合を排除する．

学習や生成は AI 橋渡しクラウド上で行い、GPU として NVIDIA Tesla V100 を用いる．

4.3　評価手法

モデルの評価として、自動的な評価と人手による評価を行う．

表 2　マルチタスク学習による応答生成の自動評価

モデル	BLEU	distinct-1	distinct-2	単語数
R	32.35	5.87	30.48	14.12
R+E6	32.29	5.93	30.48	14.12
R+E6+E2	32.39	**6.00**	**30.77**	14.11
R+E6+E12	**32.55**	5.89	30.57	**14.14**
R+E6+E2+E12	32.29	5.91	30.47	14.12

表 3　マルチタスク学習による応答生成の人手評価

モデル	感情考慮	流暢さ	有益さ	発話関連
R	3.44	3.48	3.63	3.55
R+E6	**3.59**	**3.82**	3.62	**3.96**
R+E6+E2	3.58	3.75	**3.74**	3.70
R+E6+E12	3.52	3.48	3.55	3.58
R+E6+E2+E12	**3.59**	3.75	3.57	3.64

自動評価　まず出力として得られる応答がどれだけ正解の応答と関連しているかを BLEU [17] によって評価する．そして出力される応答が語彙的に多様であるかどうかを distinct [18] で評価する．distinct に関しては、それぞれ Unigram と Bigram に注目した distinct-1 と distinct-2 を用いる．さらに出力の応答に含まれる単語数の平均も比較する．これは生成された応答が長いものであるほど、それが普遍的でないことを仮定している．

人手評価　人手による評価はクラウドソーシングで行い、プラットフォームとして Amazon Mechanical Turk を用いる．内容は EmpatheticDialogues [4] のそれにならい、評価すべき観点として感情考慮・流暢さ・有益さ・発話関連の 4 つを設定する．それぞれは生成された応答が発話の感情を考慮しているか、生成の応答が構文的に正しいものか、生成された応答が発話にとって有益な情報を提供しているか、応答の内容が発話と適切に関連しているかを問うものである．Test データから無作為に抽出された 100 サンプルを対象に、上記の 4 観点を 5 段階で評価してもらう．ワーカとしては US の在住者を指定し、各サンプルの各観点ごとに 7 ワーカを要求する．最終的なスコアは 7 ワーカによる値の平均値として得る．

ワーカに尋ねる質問の例を図 2 に示す．

4.4　結果

応答生成を R と表し、Twitter Emotion Corpus・SST-2・CrowdFlower のデータセットに関する感情認識をそれぞれ E6・E2・E12 と表す．また E6・E2・E12 のロスに対する重みをそれぞれ λ_{E6}・λ_{E2}・λ_{E12} と表す．

表 4 各モデルによる応答の生成例

発話	正解応答	R	R+E6
Yes, my parents, and soon my brother as well.	I didn't know you had a brother.	I see. In a few months, you'll be moving to a new place.	I hope I get to meet you soon.
It falls on the fifth day of the fifth lunar month.	Could you tell me the origin of the two festivals?	Oh, then on that day?	Oh? Why is that day?
How smart of you to book the tickets before hand!	Oh, do you really think so?	Sorry, I didn't think of it that way.	Yeah, yeah, I know. It was the last thing I wanted to do.
I can't. I'm waiting for Paul, but he's late.	I'll say the movie's starting in the minute.	Would you like me to call him?	Where is he?
thank you. I think I should also have a look at the Internet to see if anyone's got a sublet.	how much do are you looking to spend?	if you need a flat, you can always look online. Do you have any questions?	good idea. You can ask around at the office. If you need a flat, they'll probably have one.

表 5 ロスに重みづけを施した応答生成の自動評価

$(\lambda_{E6}, \lambda_{E2}, \lambda_{E12})$	BLEU	distinct-1	distinct-2	単語数
(1, 0, 0)	32.29	5.93	30.48	14.12
(.5, .5, 0)	32.48	5.86	30.54	**14.15**
(.5, 0, .5)	**32.52**	5.93	30.62	14.04
(.33, .33, .33)	32.43	**5.97**	**30.81**	14.01

表 6 ロスに重みづけを施した応答生成の人手評価

$(\lambda_{E6}, \lambda_{E2}, \lambda_{E12})$	感情考慮	流暢さ	有益さ	発話関連
(1, 0, 0)	3.59	3.82	3.62	**3.96**
(.5, .5, 0)	**4.00**	**4.16**	**4.01**	**3.96**
(.5, 0, .5)	3.37	3.60	3.37	3.36
(.33, .33, .33)	3.63	3.37	3.49	3.66

マルチタスク学習 自動評価の結果を表 2 に示す．表 2 からは R+E6+E2 と R+E6+E12 がそれぞれ distinct と BLEU を最大化していることがわかる．つまり提案のマルチタスク学習においては、Coarse-Grained と Fine-Grained な感情認識がそれぞれ多様性と関連性に効果的である．

人手評価の結果を表 3 に示す．感情考慮に関していえば、タスクに感情認識を含むすべてのモデルが応答生成のみによるモデルを上回るスコアとなっている．そして 4 観点のすべてを見てみると、とくに R+E6 が高いスコアをもたらしている．提案のマルチタスク学習は生成される応答をより感情が考慮されたものにするだけでなく、流暢さや発話関連といった他観点の品質も改善しうる．

得られたモデルによる生成例を表 4 に示す．R と R+E6 を対象に、与えられた発話とその応答を比較する．なお生成例については、E6 をタスクに追加したことで人手評価の感情考慮が大きく向上したサンプルを選ぶ．表 4 からは、R と比べて R+E6 に "Yeah, yeah, I know." や "good idea." といったより感情に敏感な文が含まれていることがわかる．

ロスの重みづけ 自動評価と人手評価の結果をそれぞれ表 5 と表 6 に示す．自動評価では、とくに E12 を伴うモデルで重みづけによるスコアの向上が確認できる．一方で人手評価も重みづけがいくつかのスコアを向上させており、中でも $(\lambda_{E6}, \lambda_{E2}, \lambda_{E12}) = (.5, .5, 0)$ のケースがひときわ高いスコアを出している．したがって各ロスに対する重みづけは生成される応答の品質を改善しうるもので、今回の条件においては E6 と E2 の影響を半分に抑えることがもっとも効果的である．

5 おわりに

人間とコンピュータによる対話の、ニューラルネットワークにもとづく応答生成の品質改善に取り組んだ．感情という側面に注目し、生成と分類のタスクを対象とするマルチタスク学習の応答生成モデルを提案した．自動評価と人手評価を行い、提案のモデルが感情をはじめとするいくつかの性能を向上させることを確かめた．また学習の際に計算される各ロスに重みづけを施すことで、さらなる品質の改善を図った．結果として重みづけがいくつかの評価を向上し、パラメータ更新のバランスも重要であることがわかった．

本研究は対話の感情に焦点を当て、発話の感情を考慮した応答の生成を行った．一方で本研究は応答の感情に注目しておらず、それは今後の課題である．応答がもつべき感情を推定することや、指定された感情のもと応答を生成することに取り組みたい．また今回の実験はあえて対話の文脈を一切考えていない．過去の発話やそれによる感情の変遷を生成する応答に反映させることも必要であり、それも今後取り組むべき課題である．

謝辞

本研究は JSPS 科研費 JP18H03286 の助成を受けた．

参考文献

[1] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all to-

gether: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, Online, July 2020. Association for Computational Linguistics.

[2] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2019.

[3] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] A. Kumar, A. Ekbal, D. Kawahra, and S. Kurohashi. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[8] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018.

[9] Rohola Zandie and Mohammad H. Mahoor. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*, 2020.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[11] Wei Wei, Jiayi Liu, Xianling Mao, Guibin Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. Target guided emotion aware chat machine. *arXiv preprint arXiv:2011.07432*, 2020.

[12] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Laura-Ana-Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2104–2119, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[14] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[15] Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 246–255, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[16] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[18] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.