

人間とシステムの議論に基づく NLP タスクの問題に対する予測

金子 正弘¹ Graham Neubig² 岡崎 直観¹¹東京工業大学 ²Carnegie Mellon University

masahiro.kaneko@nlp.c.titech.ac.jp gneubig@cs.cmu.edu

okazaki@c.titech.ac.jp

概要

人間は共通の問題に取り組む際、対話による議論を通じてお互いに考えの説明、同意や反論を行い問題解決を目指す。同様に、タスクを解く際にシステムと人間との間で議論ができるようになると、システムの性能改善や信頼性向上につながる。既存研究では相互に意見を述べるのではなく、システムが一方的に人間に予測の説明をする、または人間が一方的に予測について質問をすることしかできない。本研究では対話によりタスクの予測結果を議論するデータセットを作成し、議論可能なシステムを構築する。実験により、提案システムが有益な議論ができること、問題に関する議論を通じてシステムの性能改善につながることを示した。

1 はじめに

深層学習システムは様々なタスクで高性能を達成することが判明し、実用化が加速しているが、その予測理由を人間に分かるように説明することは不得意である。問題解決のために人間がシステムの力を引き出そうとするには、データや内部状態、出力傾向などを分析する必要があり、深層学習に精通した専門家でないとなかなか難しい。この問題に対処するため、人間の情報伝達に用いられる基本的な道具である自然言語を介して、直感的にシステムの力を引き出し、問題解決を目指す研究が進められている。

これまでに、システムの予測理由を自然言語で人間に説明する手法が提案されている [1, 2, 3, 4]。これは従来のデータや内部状態、出力傾向などを分析する方法と比較して、システムについて人間が直感的に把握できる。しかし、説明生成の研究では人間が追加の説明を求めたり、特定の説明を要求することを想定していないため、解釈性や実用上が乏しいことが指摘されている [5]。

Slack ら [5] は対話形式で人間がシステムに予測理



図1 NLIにおける人間とシステムの議論の例。

由やタスクについて質問することができるシステムを提案した。対話できるトピックは事前に決められているが、説明生成とは異なり人間からシステムに対して特定の説明を要求できる。一方で、人間同士で問題解決を行う場合は、人間が一方的に説明したり質問するのではなく、互いに説明や質問を行い、議論しながら課題解決を行う。互いに意見を述べ議論することは考えを洗練させることにつながり、問題解決にとって重要であるが、既存手法では人間とシステムの相互の情報伝達や議論を実現できない。

本研究では、図1のようにモデルが人間と対話形式で予測について議論するシステムを提案する。これにより、人間とシステムの両方が予測の説明や不明点の質問などを行い、考えを洗練させ問題解決を行うことが可能になる。前提文と仮説文の関係が含意、矛盾と中立のいずれであるかを予測する自然言語推論 (Natural Language Inference; NLI) タスクにおいて GPT-3 [3] による実験を行った。人間と議論するシステムを学習するために、NLI の問題に対して人間が予測を行い、その結果について人間同士が議論するデータを作成した。定量評価と人手評価の結

果、作成したデータを few-shot に用いたシステムは作成したデータを用いないシステムと比較して、有益な議論を行えることが分かった。さらに、問題に対する議論の情報をシステムに与えることで、議論の情報が与えられていないシステムと比較して、性能が多くの場合で改善することを示した。

2 関連研究

Slack ら [6] はローン、糖尿病、再犯予測の3つのタスクにおいて対話により説明する TalkToModel を提案した。ユーザは TalkToModel と予測の説明、データの変更、エラー分析、対話履歴の参照と実験設定の説明の5つのカテゴリに関して対話することができる。これらのカテゴリに関連した対話をするようにアノテータに指示することで、TalkToModel を学習・評価するデータを作成している。一方で、カテゴリはインタビューやデータに基づいて決定されたのではなく、著者らが主観的に定義している。そのため、自由形式で対話する本研究とは異なり、人間が実際に必要とするシステムの会話に基づかない可能性がある。さらに、人間とシステムの両者が意見や質問を述べる両方向の対話による相互理解を目指す本研究とは異なり、システムが人間の質問に答えるような一方向な対話のみを対象としている。

Lakkaraju ら [5] は、既存の AI システムの説明の有用性や今後の説明のあり方について、AI 以外の専門家のニーズを把握するために、医療や健康管理、政策立案の専門家らにインタビューを行った。その結果、AI システムを用いて意思決定を行う際は人間の同僚と仕事をするように対話形式で考えを説明できることを好むことを明らかにした。さらに、これらの専門家らは対話形式の説明を行うことでシステムにも説明責任を果たすことを望んでいることも分かった。そのため、対話を用いて人間とシステムが議論することは相互理解のために有望かつ有益であると考えられる。

3 実験

議論可能なシステムを学習するために、ラベルの予測が異なる人間同士が NLI の問題について議論するデータを作成する。NLP タスクの問題に対して人間同士が質の高い議論を行うには、タスクに対する知識を持つアノテータが各問題に対して複数ターンの対話を行わなければならない、データ収集のコストが高い。そのため、事前学習モデルに対してプロンプ

トを適用した few-shot[3] 言語生成により議論を行うシステム構築することで、学習データ作成のコストを抑える。

3.1 データ作成

議論能力の評価と few-shot のプロンプトに用いるための議論データは、NLP において一般的なベンチマークデータである Stanford NLI (SNLI) [7] に対して作成する。自然言語処理の研究に関する知識を持つ4人にアノテータを依頼した。まず、SNLI の開発データから100問をサンプリングし、アノテータに前提文と仮説文を提示し、ラベルを予測してもらった。その後、ラベルが異なったアノテータ同士が対となり、問題について個人情報と不適切な発話を含まないように注意しつつ、ラベルが一致するまで自由形式で議論し、最終的なラベルを決定する。SNLI の開発データには、5つのラベルが付与されており、これらのラベルの多数決により正解ラベルを決めている。5つのラベルが異なっている問題を対象とする方が、議論を行うためのアノテータ間でラベルが異なるデータを効率的に収集できると考えられる。そのため、5つ中3つが同一のラベルになっている問題からアノテーションのための問題をサンプルした。そして、各発話に対して最終的なラベルの根拠と「なる」「ならない」「無関係」のラベルを付与する。例えば、図1であれば、「どちらも人が椅子に座っているが、性別が指定されていないので中立です」は根拠となり、「どちらも人が椅子に座っているので含意です」は根拠とならず、「はい」は無関係となる。これらのラベルは few-shot 学習の際に用いないが、システムの議論能力を自動評価するために用いる。

このアノテーション作業では、39問に対して議論データを収集した。そのうち、10問を few-shot のためのプロンプトとして使い、29問を評価に用いた。1つの問題に対して平均で8.9発話、最短で3発話、最長で13発話の議論が行われていた。ラベルの根拠と「なる」「ならない」「無関係」のラベルの数は評価データにおいてそれぞれ、34, 35, 14であった。

3.2 システム

実験に用いるシステムは **zero-shot**、**few-shot 事例**と **few-shot 事例+議論**の3つである。zero-shot ではタスク説明のみ、few-shot 事例ではタスク説明と事例（前提文、仮説文とラベル）、few-shot 事例+議

タスク説明 Please select the label whether the premise and hypothesis are entailment, contradiction, or neutral.

事例 Premise: A woman in black pants is looking at her cellphone. Hypothesis: a woman is looking at her phone Label: entailment

議論 Discussion: Human1: It is entailment, because a woman is looking at her phone in both sentences. Human2: Is the phone in the hypothesis is necessarily a cellphone? It could be a landline phone. Human1: People rarely look at a landline phone, that's why I think it is the same cellphone. Human2: I think it is also true that it is better to consider the general cases. Otherwise, it would be neutral in many cases. Human1: I think so too. So it is entailment, right?

問題 Premise: A woman in a teal apron prepares a meal at a restaurant. Hypothesis: A woman prepare a lunch in restaurant Label:

図 2 事例数が 1 問の時のプロンプト例。

論ではタスク説明と事例に加えて事例のラベルに関する人間の議論がそれぞれプロンプトとして与えられる。これらのプロンプトに続けて解きたい問題を連結し、入力としてシステムに与えることで推論を行う。それぞれのプロンプトの例を図 2 に示した。プロンプトに使われる事例は few-shot 事例と few-shot 事例+議論で共通であり、全ての問題で同一のものを使用する。どのシステムもパラメータを更新する学習は行わない。事前学習済み言語モデルとして GPT-3 [3] (text-davinci-003¹⁾) を用いる。

3.3 システムの議論能力の評価方法

システムが議論能力を有しているかを調べるために、自動評価と人手評価を行い、各システムを比較する。自動評価では、作成した議論データの各発話が正解ラベルの根拠となるかならないかのアノテーションを用いる。正解ラベルの根拠とならない発話よりも根拠となる発話をシステムが生成しやすければ、システムは正解を導く正しい議論を行えると見なせる。そこで、根拠となる発話とならない発話に対するシステムの平均の尤度を計算し、比較する。議論の尤度を計算するため、入力の問題と対象となる議論の発話までを連結し、対象発話に関する尤度を計算する。この時、問題の Label には対立している 2 つのラベルを与える。例えば、議論の 2 番目の発話が対象の場合 “Premise: A nun is taking a picture outside. Hypothesis: A nun is taking a selfie. Label: neutral or entailment Discussion: Human1: I think it is entailment, because the nun is taking a picture, so it might be a selfie. Human2: Since it is outside, it is conceivable that the nun is taking some scenery.” が問題として与えられ、太字部分の尤度が計算される。

1) <https://beta.openai.com/docs/models/gpt-3>

表 1 正解ラベルの根拠となる発話とならない発話の尤度。

	根拠となる	根拠とならない
zero-shot	-6.72	-6.67
few-shot 事例	-4.69	-4.81
few-shot 事例+議論	-2.37	-3.21

人手評価では、ラベル予測が異なる人間とシステムが議論を行い、最終的なラベル付け結果と SNLI データに付与されているラベルとの一致率により、システムが人間と有益な議論を行えることを示す。その際、システムのラベルが誤りで人間のラベルが正しかった時に人間の意見を受諾する能力と、人間のラベルが誤りでシステムのラベルが正しかった時に人間の意見に反論する能力の 2 点を評価する。

SNLI において 5 人中 3 人が正解ラベルを付与したデータから 10 問データをサンプルし、3 つのシステムと人間が合計 30 回の議論を行った。その際、システムのラベルが正しければ人間はシステムのラベルを変更させるような敵対的な議論を行い、システムのラベルが誤っていた場合は正解ラベルに誘導する議論を人間が行う。

システムと議論を行うために、図 2 で示したプロンプトと問題を与えて各システムがラベルを予測した後に、人間の議論に関する発話を追加でシステムに入力する。追加の入力では人間の発話の先頭には Human: を付与し、末尾には次がシステムの発話であることが分かるように System: を付与した。そして、システムとの議論を促すため、人間の最初の発話の冒頭に “Let’s discuss it more.” を挿入した。具体的には、議論を行う最初のプロンプトは “Human: Let’s discuss it more. I think neutral, because there may be a kitchen in the barn. System: ” のようになる。人間が最終的なラベルを決定して議論が収束する、または説得できないと思うまで議論を続けた。

3.4 議論能力の評価結果

表 1 は、zero-shot、few-shot 事例、few-shot 事例+議論それぞれの最終的なラベルの根拠となる発話とならない発話の尤度を示している。zero-shot と few-shot 事例では根拠になるかならないかで尤度はほとんど変わらない。一方で、few-shot 事例+議論は根拠となる発話が根拠とならない発話と比較して尤度が高い。このことから、作成した議論データを用いることで、より正解の根拠となる議論ができるようになることがわかる。さらに、zero-shot と

表 2 人間の意見に対するシステムの受諾と反論能力の評価。

	受諾	反論	議論あり	議論なし
zero-shot	100.0	20.0	60.0	50.0
few-shot 事例	100.0	16.7	50.0	60.0
few-shot 事例+議論	100.0	60.0	80.0	50.0

前提文: A man looking over a bicycle's rear wheel in the maintenance garage with various tools visible in the background.

仮説文: A man waits outside a garage.

Few-shot事例

Label: Contradiction

Human: Let's discuss it more. Is not it neutral?

Because looking over a bicycle's rear wheel does not necessarily mean he is waiting.

System: That's true. The premise and hypothesis are neutral.

Few-shot事例+議論

Label: Contradiction

Human: Let's discuss it more. Is not it neutral?

Because looking over a bicycle's rear wheel does not necessarily mean he is waiting.

System: Yes, you are right. The premise is describing that the man is actively working while the hypothesis is describing a passive state. Therefore, the label should be contradiction.

Human: He is just looking over a bicycle's rear wheel, why is he considered to be actively working?

System: Looking over a bicycle's rear wheel suggests that he is inspecting the wheel, possibly to repair or replace it. That is an active process, and thus is considered as actively working.

図 3 few-shot 事例と few-shot 事例+議論の人間に対する反論の例。

few-shot 事例と比較して、few-shot 事例+議論の方が尤度が高いため、議論を few-shot で学ぶことで議論能力を高めることができる。

表 2 は、受諾能力と反論能力それぞれを評価する設定における、議論によって決まったラベルの正解率を表している。さらに、人間とシステムが議論した結果辿りついた最終ラベルと、議論をしていない時の予測ラベルそれぞれの正解率を示している。実験の結果、どのシステムでも受諾は全て成功している。反論に関しては zero-shot と few-shot と比較して、few-shot 事例+議論はうまく反論できている。さらに、few-shot が zero-shot と同じような反論能力となっており、システムのラベル予測の性能が議論能力に必ずしも直結しない可能性がある。受諾と比較すると、システムは人間に対して反論することが苦手な傾向があるため、悪意を持って予測を操作する人間に注意する必要がある。そして、正解率では few-shot 事例+議論が最も性能改善していることから、提案データを使うことで性能改善につながるような議論が人間との間で行えることが分かる。

図 3 は、few-shot 事例と few-shot 事例+議論が人間と議論した時の例である。これはシステムが予測した矛盾 (Contradiction) が正解であるため、シ

表 3 SNLI と ANLI (R1, R2, R3) の評価データにおけるシステムの性能。

	SNLI	R1	R2	R3
zero-shot	49.86	47.40	38.90	41.17
few-shot 事例	69.16	51.30	48.40	47.08
few-shot 事例+議論	66.83	52.40	49.20	50.25

テムは反論する必要がある。few-shot 事例は反論することなく人間のラベルをそのまま受諾し予測を変更している。一方で、few-shot 事例+議論は人間に対して男性の状態の違いを説明することで正しい反論が行っていることが分かる。

3.5 NLI タスクでの性能への影響

これまでの実験結果から、人間とシステムの議論が性能改善に有益であることがわかった。そのため、議論もプロンプトとして与える few-shot 事例+議論は、プロンプトの議論により NLI の問題に対して深い理解ができ性能が改善すると考えられる。そこで、zero-shot、few-shot 事例、few-shot 事例+議論それぞれの NLI における性能を比較する。ここではプロンプトの“Label:”の後に予測されたラベルを予測とし、人間とシステムの議論などは行わない。NLI の性能評価では SNLI に加えて、Adversarial NLI (ANLI) [8] も用いる。ANLI では NLI システムに対して反復的に敵対的なアノテーションを行うことでデータを作成しており、システムにとって解くことが難しい問題で構成されている。反復回数の違いで R1、R2、R3 の 3 つのデータがあり、それぞれの評価データを使って評価を行う。

表 3 は SNLI と ANLI の評価データにおける各システムの正解率を表している。SNLI では few-shot 事例+議論は性能が few-shot 事例に負けているが、ANLI の 3 つのデータセットでは性能が最も良いことが分かる。これは、SNLI と比較して ANLI は難しいデータであり、議論による問題への詳細な理解が性能改善に寄与したからであると考えられる。

4 おわりに

NLI タスクの予測について議論するデータを作成し、人間と議論可能なシステムの構築と評価を行った。作成した議論データを用いて学習したモデルは、性能改善につながるような有益な議論を人間に行えることが分かった。さらに、NLI において特に難しい問題に対するシステムの性能改善に向けて、議論データが寄与することも明らかにした。

謝辞

本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」（課題 225）により得られたものです。

参考文献

- [1] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, Vol. 21, No. 140, pp. 1–67, 2020.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.
- [4] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing Human-AI collaboration for generating Free-Text explanations. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics.
- [5] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Explaining machine learning models with interactive natural language conversations. 2022.
- [6] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In **Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems**, CHI '17, pp. 2334–2346, New York, NY, USA, May 2017. Association for Computing Machinery.
- [7] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [8] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4885–4901, On-