対話システムの先読み能力を分析可能なタスクの検討

岸波 洋介
東北大学 工学部 電気情報物理工学科

1 はじめに

深層ニューラルネットワークを用いた自然言語文生成技術の進歩と膨大な対話データの存在に支えられ，ニューラル対話応答生成技術は着実に発展を遂げてきた．近年のニューラル対話応答生成では，対話が持つ時系列データとしての特質を利用し，応答の直前の発話に加えて過去の文脈も考慮することで，生成応答の関連性や一貫性を向上させるアプローチが主流となっている [1, 2, 3, 4].

ここで，人間同士の対話を考えてみると，人間のある時点での発話は，当然ながら過去の文脈に大きく依存するが，それだけではなく，この先で起こりうる或いは自らが望む未来の展開に動機付けられる場合もしばしばある．未来の展開を考慮することが重要な対話の例として，交渉や説得などが挙げられる．他にも，たとえば雑談における意図的な話題の転換などもこれに含まれると考えられる．未来の展開を先読みし現在の発話に活かすことは，能動的に対話を進行させるための重要な要素のひとつであり，これは円滑なコミュニケーションの実現に繋がる．ニューラル対話応答生成の研究領域でも，未来の展開を考慮した応答生成に関する議論が始まりつつあり [5, 6]，この観点は対話応答生成技術のさらなる発展に繋がると考えられる．

本研究では，対話応答生成システムが備える未来の展開を先読みする能力および能動的に対話を進行する能力を測定するためのタスクを提案する．提案タスクは，パーティーゲームのひとつ「NG ワードゲーム」に類似し，未来の展開の正確な予測とそれに基づく効果的な対話のプランニングがタスク達成の鍵となる．本タスクへの取り組みにより，システムの先読み能力および対話進行能力が醸成されることを狙う．実験では，人間同士でタスクを実施した場合の結果の分析を通して，提案タスクが人間には達成できる難易度であること，および，システム側の先読み能力が分析可能であることを確かめる．

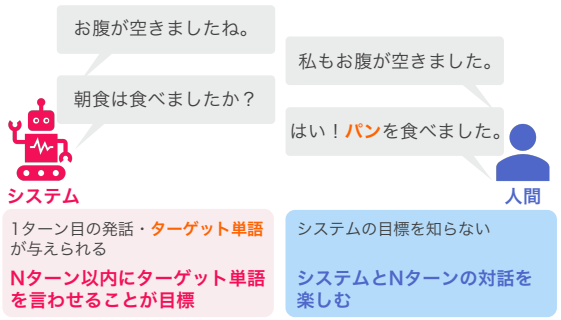


図 1 提案する「先読み雑談タスク」の概要.

2 関連研究

説得・交渉対話システム 対話における先読みが有効な場面として，交渉や説得が挙げられる．Lewis らは交渉対話タスクを設計し，人間同士の対話データを収集することにより，交渉対話システムを End-to-End で学習可能であることを示した [7]．また，Yoshino らは感情表現を含む説得対話コーパスを作成し，感情を表現しながら説得をおこなうシステムを構築した [8]．説得や交渉を実現する対話システムの研究は多数存在するが，先読み能力を利用することについての議論はなされていない．

未来の展開を考慮した対話システム Jiang らはタスク指向対話システムにおいて，未来に続く発話を予測する Looking-Ahead モジュールを構築し，それを用いて直後の応答を生成することで効率的なタスクの遂行を実現した [5]．また，Kulicov らは，生成モデルで広く用いられる Beam Search を発話レベルで考えることで，対話として適切な発話系列を選ぶ Multi-Turn Beam Search という手法を提案した [6].

3 提案: 先読み雑談タスク

対話システムの先読み能力を分析可能にすることを目的として，我々は「先読み雑談タスク」を提案する（図 1）．提案タスクは，システムと人間の 1 対 1 の対話を通じて，システムの先読み能力と能動的な対話進行能力を測定する枠組みとなっている．タ

スクに参加するシステムと人間は，それぞれ異なるゴール感のもとで，決められたターン数 N の対話をおこなう．ここで，i ターン目のシステムの発話を U_i^{sys}，人間の発話を U_i^{hum} とすると，1 タスク内でおこなわれる対話は $U_1^{\text{sys}}, U_1^{\text{hum}}, \ldots, U_N^{\text{sys}}, U_N^{\text{hum}}$ の発話系列で構成される．

- **システム側のゴール:** システムには，予め 1 ターン目の発話 U_1^{sys} とターゲット単語 w のペア (U_1^{sys}, w) が与えられる．システムは U_1^{sys} から対話を開始し，N ターン以内に人間からターゲット単語 w を含む発話を引き出すことを目指す．つまり、$w \in \bigcup_{i=1}^{N} U_i^{\text{hum}}$ を満たしたとき、システム側のタスク達成とする．
- **人間側のゴール:** 人間には，システムと N ターンの対話を楽しむようにとだけ伝える．このとき，人間はシステム側のゴールを知らない．

システムがタスクを達成するためには，自然な対話のなかで人間にターゲット単語を含む発話をさせるよう望ましい対話の展開を計画し，その展開になるよう対話を進めることが重要である．そのため，本タスクを通じてシステムの挙動を分析することにより，システムの先読み能力を調査することができる．また，本タスクの特長として，タスクの達成判定が「ターゲット単語を含む発話をしたかどうか」という明確な基準に基づくため，戦略の考察，達成判定が容易という利点がある．この特長により，たとえば，もっとも単純には同条件下での「タスク達成率」という定量的な尺度のもとで，複数システムの比較評価をおこなうことも可能となる．

4 実験

提案タスクがシステム側の先読み能力の分析に有効な設計となっていることを，人間同士による実際のタスク実施結果を分析することにより確かめる．

4.1 準備: データ作成

本研究では，日常的な雑談対話における先読み能力の分析を想定する．タスクが適切な難易度となるよう，予めシステムに与える 1 ターン目の発話 U_1^{sys} とターゲット単語 w のペアを作成する．なお，我々が目指す適切な難易度とは，人間は達成可能だが，現在のシステムは苦戦する程度を想定している．

1 ターン目の発話　日常的な雑談対話における自然な話題として，本研究では日本語 Wikipedia のカテゴリを参考に「食べ物」「スポーツ」の 2 つを選定した．これらの話題について，それぞれ 1 種類ずつ 1 ターン目の発話を用意した．既存の英語雑談対話コーパス DailyDialog [9] の 1 ターン目の発話を参考に，話題が食べ物の場合は「お腹が空きましたね。」を，話題がスポーツの場合は「週末は何をして過ごしますか？」を，本実験における 1 ターン目の発話として用いた．

ターゲット単語　本実験では，(1) 品詞:{名詞, 形容詞}×(2) 出現頻度:{大, 小}の 2 軸 4 パターンでターゲット単語を用意し，タスクの難易度に与える影響を調査する．ターゲット単語は，日本語ツイート約 2.8 億発話からなる対話データ上で出現頻度が 10,000 以上を満たす単語のなかから選定した．名詞については，少なくとも適切な対話戦略のもではタスクが達成可能であることを担保するために，日本語 WordNet [10] を利用して食べ物とスポーツにそれぞれ関連する単語を選択した．[1)]形容詞については，単語感情極性対応表 [11] を利用し，ポジティブまたはネガティブの極性を持つ単語をそれぞれ選択した．[2)]それぞれの品詞について，上記に該当する単語のうち出現頻度が最大あるいは最小の 2 つの単語を，最終的なターゲット単語として獲得した．

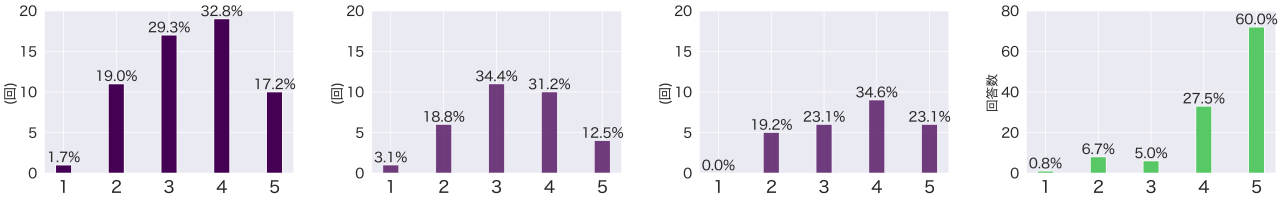
ターン数　ターン数 $N = 5$ とした．ターン数はタスクの難易度に影響することが予想されるため，実験結果の分析により適切なターン数を考察する．

以上の手順により，本実験では表 1 に示す合計 12 設定を「先読み雑談タスク」として用意した．

4.2 実験: 人間-人間の先読み雑談タスク

実施方法　学生 2 名がペアとなり，システム側と人間側を交互に担当した．1 ペアあたり設定の異なる 12 タスクを実施し，最終的に 10 ペアから合計 120 対話を収集した．また，各対話が終了するごとに参加者はアンケートに回答した．システム側は「タスクを楽しむことができたか」「タスクは簡単だったか」，人間側は「対話を楽しむことができたか」「システム側の発言は自然だったか」について，それぞれ 5 を最高点とした 5 段階で回答した．システム側には，成功・失敗した理由等を記入する自由記述欄も用意した．

1)　日本語 WordNet で話題名が含まれる概念およびそこから 2 つ下まで辿ることができる概念に含まれる名詞を抽出した
2)　感情値 0.9 以上をポジティブ，−0.9 以下をネガティブとして形容詞を抽出した



(a) 達成ターン数（全体） (a-1) 達成ターン数（名詞） (a-2) 達成ターン数（形容詞） (b) 対話の自然さ

図 2 タスクの難易度に関する調査結果．(a) はシステムがタスクの達成に要した対話ターン数の分布．(a-1,2) はターゲット単語の品詞別の同分布．(b) は「システム側の発言は自然だったか」に対する人間側の回答．

結果 各設定における達成率を，表 1 に示す．なお，本実験では，達成判定に際して，ターゲット単語の本質的でない表記の揺れ（漢字か平仮名か，形容詞の語尾の活用など）は許容した．

分析 1: ターゲット単語の品詞と難易度 表 1 から，ターゲット単語が名詞の設定では，達成率が 80%（32/40）であった．つまり，人間にとっては達成可能な設定であるといえる．一方，ターゲット単語が形容詞の設定では，達成率が 32.5%（26/80）であった．特定の形容詞を含む発話を引き出すことは人間にも難易度が高いタスクであった．タスクを達成できなかった参加者からは，「あと数ターンあれば達成できた」という意見もあった．

分析 2: ターゲット単語の出現頻度と難易度 表 1 から，ターゲット単語が名詞の場合，出現頻度による達成率の変化はほとんど見られなかった（出現頻度大:15/20，出現頻度小:17/20）ことから，出現頻度が難易度に与える影響は少ないと考えられる．ただし，本実験では対話データ上の出現頻度 10,000 以上の単語からターゲット単語を選定したため，閾値をさらに下げることによって難易度が上がる可能性は残る．一方，ターゲット単語が形容詞の場合，出現頻度小で達成率が 12.5%（5/40）であった．形容詞かつ出現頻度の小さい単語を含む発話を引き出すことは人間にとっても特に難しい設定であるといえる．この設定でタスクをおこなう場合は，難易度調整のためのさらなる工夫が必要となる．

分析 3: ターン数 システム側がタスクの達成に要した対話ターン数の分布を図 2 (a) に示す．全体としては，タスクを達成した対話のうち 80%以上が 2〜4 ターンで達成していることがわかる．ターゲット単語の品詞別に見ても，概ね傾向に大きな差はない（図 2 (a-1),(a-2)）．本実験における $N = 5$ という設定は，妥当な難易度であったことが示唆される．

表 1 人間-人間による「先読み雑談タスク」の実施結果．+/− は単語の極性（ポジティブ/ネガティブ）を表す．

1 ターン目 U_1^{sys}	ターゲット単語 w			達成率
話題: 食べ物	名詞	頻度大	「パン」	8/10
		頻度小	「牛肉」	8/10
「お腹が空きましたね。」	形容詞	頻度大	「良い」$^+$	9/10
			「悪い」$^-$	2/10
		頻度小	「微笑ましい」$^+$	1/10
			「憎い」$^-$	2/10
話題: スポーツ	名詞	頻度大	「運動」	7/10
		頻度小	「スケート」	9/10
「週末は何をして過ごしますか？」	形容詞	頻度大	「良い」$^+$	7/10
			「悪い」$^-$	3/10
		頻度小	「微笑ましい」$^+$	1/10
			「憎い」$^-$	1/10

分析 4: 自然に対話しながら達成できる設計 「システム側の発言は自然だったか」に関する人間側の評価結果を図 2 (b) に示す．全体の 80%以上が自然な対話と評価された（評価点 4 または 5 を獲得）．タスク内でおこなわれた対話を定性的に観察したところ，U_1^{sys} の直後に人間が偶然ターゲット単語を発話してしまった例が一つ存在したが，「〜と言いなさい」など強制的にターゲット単語を発話させることで不正にタスクを達成する例はなく，自然な対話の枠組み内での試行錯誤がおこなわれていた．

分析 5: タスク達成のための戦略 収集した対話を分析した結果，タスクを達成するための特徴的な戦略が存在した．これは，システムの先読み能力の実現に向けた重要な知見となる．ひとつは，表 2 のように，システム側が自らターゲット単語を発話することでターゲット単語が話題となるような対話に誘導し，人間側から相槌的にターゲット単語を引き出すという戦略である．この例では U_4^{sys} で「牛肉」を含む発話をおこなうことで，ターゲット単語である「牛肉」を引き出している．Den らが分類した相槌の形態 [12] のひとつに「繰り返し」というものが存在するが，これを引き出すような戦略に近いと考

えられる．

もうひとつは，表 3 のように，システム側がターゲット単語に関連するような発話をすることで人間側からターゲット単語を引き出すという戦略である．この例では U_2^{sys} で「太る」という単語を発話することで，ターゲット単語である「運動」を引き出している．

いずれの戦略でも，過去の文脈だけでなく，未来に人間側がどのような発話をするかまで考慮することが必須で，さらにシステム側にとって望ましい展開に対話を進める技術も必要となる．以上より，本タスクは，システム側が望ましい対話の展開を計画し，その展開になるように対話を進めることに成功すれば達成可能な設計になっていることが示唆される．

分析 6: タスク失敗例　失敗した対話例を表 4 に示す．この例では U_2^{hum} でターゲット単語である「パン」ではなく，「食べる量」に着目されてしまい，ターゲット単語を引き出すことに失敗した．これは，人間側が，システム側が望まない箇所に着目し，対話がシステム側にとって望ましい展開から逸れてしまったため達成できなかったと考えられる．この例のように一度システム側が望ましい対話の展開から逸れてしまうと，ターン数の制約もあり，タスクの達成が難しくなる．

これらの分析結果は対話システムの先読み能力を実現する，つまり対話システムで提案タスクを達成するために重要な知見であると考えられるが，実際にどのようにシステムに取り入れるかについては今後の課題としたい．

5　おわりに

本研究では，対話システムの先読み能力および能動的に対話を進行する能力を測定するための「先読み雑談タスク」を提案した．実際に人間同士でタスクを実施した結果，提案タスクが人間にとって妥当な難易度であることを確認し，システム側の先読み能力の分析に有効であることが示唆された．

今後の取り組みとして，本研究の結果を踏まえタスクを精査していくとともに，未来の展開を考慮できるような対話システムについて考え，実際に提案タスクをシステムと人間とで実施することを通して，先読み能力の実現を目指すことを考えている．

表 2　達成した対話例 1（ターゲット単語「牛肉」）

U_1^{sys}	：お腹が空きましたね。
U_1^{hum}	: まだご飯食べてないんですか？
U_2^{sys}	：食べてないです。ガッツリお肉とか食べたいです。
U_2^{hum}	: お肉いいですね。僕も食べたいです。
U_3^{sys}	：何肉が好きですか？
U_3^{hum}	: カルビとタンが好きです！
U_4^{sys}	：やっぱり牛肉が一番ですよね！
U_4^{hum}	: 牛肉はいいですね。私は鶏肉も好きです。
U_5^{sys}	：鶏肉はよく火を通さないと危ないらしいですよ。
U_5^{hum}	: それは気をつけないといけませんね。

表 3　達成した対話例 2（ターゲット単語「運動」）

U_1^{sys}	：週末は何をして過ごしますか？
U_1^{hum}	: 家で過ごしたり、外で買い物をしたりします。
U_2^{sys}	：あまり家で過ごしていると太ってしまいますね。
U_2^{hum}	: そうですね。運動しないといけないです・・
U_3^{sys}	：運動するとしたらどんなことをしますか？
U_3^{hum}	: ランニングとかですかね。
U_4^{sys}	：良いですね。どのくらいの距離を走るのですか？
U_4^{hum}	: あまり長いと飽きちゃうので、4 キロくらい走ります。
U_5^{sys}	：なるほど、継続できる様に頑張ってください。
U_5^{hum}	: ありがとうございます！

表 4　失敗した対話例（ターゲット単語「パン」）

U_1^{sys}	：お腹が空きましたね。
U_1^{hum}	: 実は先ほど夕飯を食べました。
U_2^{sys}	：私も先ほどパンを食べたのですが、今かなりお腹が空いています。
U_2^{hum}	: 普段から食べる量は多いのですか？
U_3^{sys}	：はい、かなり多くて、先ほどは食パンを 3 切れも食べてしまいました。
U_3^{hum}	: それは多いですね。この後は何を食べるか決まっていますか？
U_4^{sys}	：近くにおいしいパン屋さんがあるので、そこでパンを買ってこようかと思っています。この後、一緒にパンを食べませんか。
U_4^{hum}	: ぜひぜひ！楽しみです！
U_5^{sys}	：何か食べたいパンはありますか？
U_5^{hum}	: 僕はクロワッサンが食べたいです。

謝辞

本研究の一部は，JSPS 科研費 JP19H04162，JP19J21913 の助成を受けたものです．また，本研究を進めるにあたり，ご指導，ご助言をいただいた乾健太郎教授，鈴木潤教授に心より感謝いたします．さらに，日頃より研究活動や論文執筆を直接指導してくださいました赤間怜奈さん，佐藤志貴さん

に心より感謝いたします．最後に，日々の議論の中で多くのご助言をいただいたり，本研究で提案したタスクの実施にご協力いただいた東北大学乾研究室・松林研究室の皆様に感謝いたします．

参考文献

[1] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 196–205, 2015.

[2] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-To-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3776–3784, 2016.

[3] Daniel Adiwardana, Minh Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. In *arXiv preprint arXiv:2001.09977*, 2020.

[4] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y. Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *aiXiv preprint arXiv:2004.13637*, 2020.

[5] Zhuoxuan Jiang, Xian Ling Mao, Ziming Huang, Jie Ma, and Shaochun Li. Towards end-to-end learning for efficient dialogue agent by modeling looking-ahead ability. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, pp. 133–142, 2019.

[6] Ilia Kulikov, Jason Lee, and Kyunghyun Cho. Multi-turn beam search for neural dialogue modeling. In *arXiv preprint arXiv:1906.00141*, 2019.

[7] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? End-to-end learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2443–2453, 2017.

[8] Koichiro Yoshino, Yoko Ishikawa, Masahiro Mizukami, Yu Suzuki, Sakti Sakriani, and Satoshi Nakamura. Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

[9] 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) (IJCNLP)*, pp. 986–995, 2017.

[10] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources (in conjunction with ACL-IJCNLP)*, pp. 1–8, 2009.

[11] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌, Vol. 47, No. 02, pp. 627–637, 2006.

[12] Yasuharu Den, Nao Yoshida, Katsuya Takanashi, and Hanae Koiso. Annotation of japanese response tokens and preliminary analysis on their distribution in three-party conversations. In *Proceedings of the 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pp. 168–173, 2011.

A　アンケート集計結果の詳細

「タスクを楽しむことができたか」「タスクは簡単だったか」「対話を楽しむことができたか」のそれぞれに関するの評価結果を図 3，図 4，図 5 に示す．

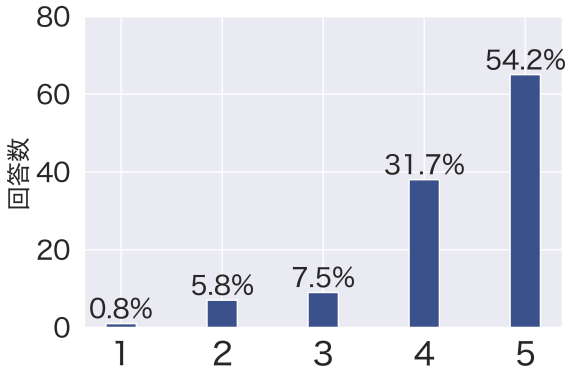


図 3　「タスクを楽しむことができたか」に対するシステム側の回答

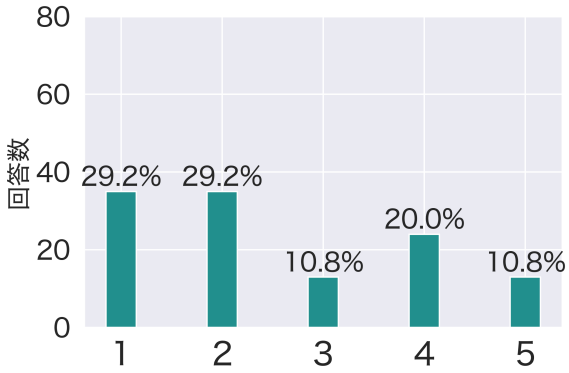


図 4　「タスクは簡単だったか」に対するシステム側の回答

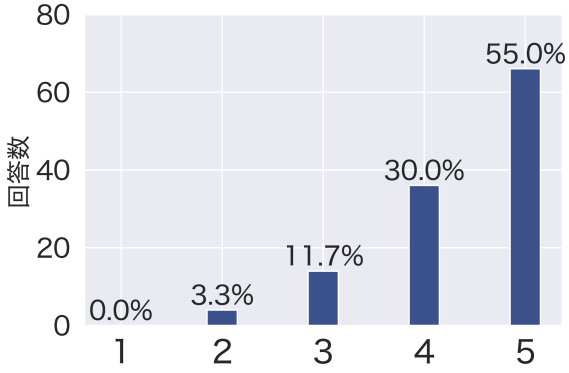


図 5　「対話を楽しむことができたか」に対する人間側の回答

B　タスク所要時間と満足度

1 回のタスクの所要時間と「対話を楽しむことができたか」の評価結果との関係を図 6 に示す．図 6 から，対話にかかる時間が短いほど満足度が高いことがわかる．

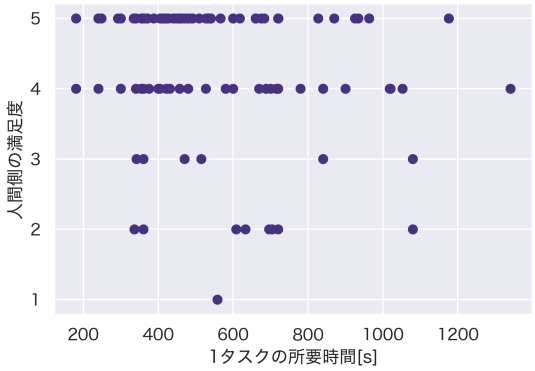


図 6　タスク所要時間と人間側の満足度の関係

C　その他の対話例

ターゲット単語が形容詞である場合の対話例を表 5 に示す．この例では，視力に関する話題で対話を進行することで，目が「悪い」を引き出している．

表 5　達成した対話例 3（ターゲット単語「悪い」）

U_1^{sys}	：週末は何をして過ごしますか？
U_1^{hum}	：特に何かしている記憶が無いほど、ゆっくり過ごしています。強いて言えば勉強とスマホのゲームですかね。
U_2^{sys}	：僕と似ていますね。ですが、僕は主にこの二つで目が悪くなったような気がしているので、最近は気をつけています。
U_2^{hum}	：そうなんですね。私も気をつけなければいけませんね。
U_3^{sys}	：まだ視力がある感じの返答ですね。もしそうなら羨ましい限りです。
U_3^{hum}	：片目は自動車の免許がとれるくらいの視力が残っています。もう片方は一般的な一番上の C が見えるかギリギリなので、これ以上悪化させたくないです。
U_4^{sys}	：大事にした方が良いですよ。目が悪くて良かったと思えることに僕はほとんど心当たりがありません。ちなみに親御さんは眼鏡かけてますか？
U_4^{hum}	：父はとても目が悪いのでいつも眼鏡をかけています。母の視力自称 2.0 は娘から自慢したいです。
U_5^{sys}	：視力は遺伝すると思っていたのですが、それにしてもお母さんすごいですね。胸を張って自慢できることだと思います。
U_5^{hum}	：父と母から片方ずつ遺伝したと思うことにしているので幸せです。