大規模言語モデルを用いた対話システムの語彙レベル制御

TSENG YI KAI[1]　徳永健伸[1]　横野光[2]
[1] 東京工業大学 [2] 明星大学
tseng.y.ab@m.titech.ac.jp　take@c.titech.ac.jp　hikaru.yokono@meisei-u.ac.jp

概要

人間同士の対話では alignment が発生することが昔からよく知られており，我々の先行研究では Lexical Level Alignment (LLA) という発話で使われる単語の語彙レベルでの alignment を提案した．本論文では，対話システムで LLA を実現するために，ChatGPT を発話生成のモデルとして採用し，生成される発話の語彙レベルを制御するための指定の語彙レベルに応じた発話生成用単語リストを作成する外部モジュールを提案する．

1　はじめに

人間同士の対話では様々な linguistic level で alignment が発生することが昔からよく知られている [Pickering and Garrod, 2006]．我々の先行研究 [Tseng et al., 2023] では，Lexical Level Alignment (LLA) という発話で使われる単語の語彙レベルでの alignment を提案した．子供や非母語話者と話す時に相手が理解しやすいようにより簡単な単語を使うことがあるが，これは LLA の一つの例である．

本研究では LLA を考慮した対話システムの実現を目指している．発話生成にはプロンプトベースの大規模言語モデルである ChatGPT [1)] を採用するが，予備実験 (付録 A) において，ChatGPT の生成する発話の語彙レベルをプロンプトのみで制御することは不十分であることが明らかになった．

本論文ではこの問題の解決に向けて，これまでの対話履歴を考慮した，指定した語彙レベルに応じた発話生成用単語リストを作成して発話生成を行う ChatGPT に提示する外部モジュールを提案し，指定した語彙レベルより難しい単語をできる限り使わずに発話を生成することを目指す．

2　関連研究

Ehara [2023] は英語学習のための文章を ChatGPT を用いて生成するシステムを開発している．システムは特に単語の習得に注目し，ユーザが選択したトピックとユーザの学習状況に応じて習得する単語を選択し，生成される文章に選択された単語が含まれるように ChatGPT に指示する．

Landwehr et al. [2023] は ChatGPT に外部のモジュールを導入し，バーチャル AI キャラクターのための記憶機能を実現している．提案システムは発話履歴から記憶のデータベースを構築し，発話を生成する度にデータベースから関連する記憶を抽出して ChatGPT へのプロンプトに挿入している．

本論文の提案システムは Landwehr et al. [2023] の構成を参考にして外部モジュールを導入して外部データベースとして分類語彙表 [National Institute for Japanese Language and Linguistics, 2004] を参照して発話生成用単語リストを作成する．そして，Ehara [2023] の手法を参考にして発話生成用単語を使うように発話生成の ChatGPT に指示する．

3　提案手法

発話生成用単語リストの作成において，最も単純な手法として指定した語彙レベルの単語を分類語彙表から選択することが考えられるが，発話生成用単語リストが極端に大きくなり，対話のトピックに関連しない単語を多く含めてしまうため，適切な返答を生成できない可能性がある．

そこで，対話の一貫性を保つために，提案手法では対話のトピックに関連する単語で，かつ，指定の語彙レベルを満たす単語のみを選択する．

具体的には，まず，ChatGPT を用いて，対話のトピックを代表するキーワードを発話履歴から抽出する (3.1 節)．次に，分類語彙表の分類情報を参照して，対話のトピックを代表するキーワードと同じ分類の単語を抽出し，その中で指定した語彙レベルより簡単な単語を選択して発話生成用単語リストを作成する (3.2 節)．最後に，作成した発話生成用単語リストと発話履歴を含めたプロンプトを ChatGPT に

1）https://openai.com/blog/chatgpt

提示し，それらの単語をできる限り使うように応答の発話を生成させる (3.3 節).

本論文では，我々の先行研究でも用いた分類語彙表の「聞く」の単語親密度 [Asahara, 2019] を語彙レベルとして用いる．従って，単語親密度が高いと簡単な語であるとみなす．

3.1 トピックキーワードの抽出

トピックキーワードは ChatGPT を用いて発話履歴から抽出する．

ChatGPT に提示するプロンプトは二つのメッセージから構成される (図 1)．ここで，二つ目のメッセージには直近の 20 発話を入れ，最初の発話者を “A”，二番目の発話者を “B” とする．

(System)
A と B の対話において、対話のトピックを代表する単語を抽出してください。一行に一単語を出力してください。

(User)
A: { 話者 A の発話 1}
B: { 話者 B の発話 1}
A: { 話者 A の発話 2}
...
B: { 話者 B の発話 10}

図 1 トピックキーワード抽出のプロンプト

モデルの温度パラメータは再現性のために 0 に設定する．

3.2 発話生成用単語リストの作成

発話生成用単語リストを作成するために，分類語彙表からトピックキーワードと同じ分類の単語を抽出し，候補単語集合を作成する．分類語彙表では各単語に品詞を表す一桁の整数部分と，上位語関係が反映された意味を表す四桁の小数部分から構成される分類コード[2] が付与されている．トピックに関連する単語は品詞を問わず該当するため，本論文では分類語彙表の分類コードは意味を表す小数部分のみ考慮する．

次に，ChatGPT が生成する応答の語彙レベルを制御するために，指定の語彙レベルより難しい単語を候補単語集合から除外する．最後に，最大 N 個の単語を候補単語集合からランダムにサンプリングすることで，固定サイズの発話生成用単語リストを作成する．

2) 例えば，分類コードが “2.5170” の単語は動詞であり，“自然-物質-熱” という分類に属するが，小数部分の “.5“は “自然“という部門，“.51“はその中の “物質 “という中項目，“.5170“はその中の “熱 “という分類項目を表す．

3.3 発話の生成

対話システムで用いるベースラインの ChatGPT モデルにおいて，プロンプトは状況を指定するシステムメッセージと発話履歴から構成される (図 2)．提案手法は発話生成用単語リストを提示するため

(System) あなたはユーザと雑談しています。
(Assistant) { 話者 A の発話 1}
(User) { 話者 B の発話 1}
(Assistant) { 話者 A の発話 2}
...
(User) { 話者 B の発話 10}

図 2 ベースラインモデル発話生成のプロンプト

(System)
ユーザとの対話の一貫性を保ちつつ、次の単語をできるだけ使って二文以下で応答してください。

{ 発話生成用単語 1}
...
{ 発話生成用単語 N}

(System) 必ずユーザとの対話の一貫性を保って、二文以下で応答してください。

図 3 提案システムの追加プロンプト

に，ベースラインモデル向けのプロンプトに二つのシステムメッセージを追加する (図 3)．

モデルの温度パラメータは再現性のために 0 に設定し，frequency penalty は同じトークンを繰り返す意味不明な出力を防ぐために 0.3 に設定する．

4 評価実験

提案手法によって ChatGPT が生成する発話の語彙レベルが制御できるかの評価を行った．

4.1 対話データ

評価実験で用いる対話データとして我々の先行研究でも用いた BTSJ 1000 人日本語自然会話コーパス [USAMI, 2023] の雑談対話を用いる．368 雑談対話のうち，40 ターン未満の 8 対話を除外し，残りの 360 対話において各対話の前半の最後の 20 発話を抽出してデータセットとする．抽出した 20 発話において，次の発話が対話システムのアシスタントのターンになるように，一番目の話者をアシスタント，二番目の話者をユーザと仮定する．従って，20

発話の最後のユーザの発話に対してアシスタントが応答を生成する．

4.2 設定

評価実験では指定語彙レベル LL を $0.6, 0.8, 1.0, 1.2, 1.4$ に設定し，発話生成用単語リストのサイズ N を $25, 50, 100, 200, 400, 800, 1600$ に設定する．発話生成用単語リストはランダムに生成されるため，評価実験では各対話に対して発話生成用単語リストを 10 個生成して ChatGPT に応答の発話を生成させる．従って，各 (N, LL) ペアに対して合計 3,600 個の発話が生成される．

ベースラインモデルでは，プロンプトは対話毎に一意に決まるため，各対話に対して 1 個の発話のみを生成する．

実験では，ChatGPT は “gpt-3.5-turbo-16k-0613” のモデルを用い，コンテンツフィルタは有害コンテンツと誤判断されて生成を拒否される確率を減らすように最も緩い設定にする．

5 結果と考察

5.1 無効な出力

コンテンツフィルタを最も緩い設定にしたにも関わらず，ChatGPT は入力したプロンプトと生成したテキストを潜在的に有害なコンテンツと判定して生成を拒否することがある．また，エラーメッセージや非常に長い文など，ChatGPT が想定している形式での応答ではない出力を生成することがある．本実験では，これらのような応答は無効な出力と見なした．

各 (N, LL) ペアにおいて合計 3,600 個の応答が生成されたが，その中で無効な出力の数を表 1 に示す．N が大きい場合と LL が低い場合で，ChatGPT が無効な出力を生成する傾向が見られた．このような場合，発話生成用単語リストに一般的にあまり使われない単語が出現する確率が上がるため，そのような単語が多すぎると ChatGPT は正常に発話を生成できなくなると考えられる．

表 1 無効な出力の数 (総数=3,600)

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
25	494 (14%)	396 (11%)	324 (9%)	296 (8%)	242 (7%)
50	410 (11%)	373 (10%)	296 (8%)	285 (8%)	195 (5%)
100	427 (12%)	359 (10%)	285 (8%)	267 (7%)	198 (6%)
200	622 (17%)	460 (13%)	332 (9%)	273 (8%)	203 (6%)
400	816 (23%)	581 (16%)	424 (12%)	306 (8%)	206 (6%)
800	836 (23%)	617 (17%)	428 (12%)	292 (8%)	189 (5%)
1600	812 (23%)	627 (17%)	417 (12%)	300 (8%)	203 (6%)

5.2 語彙レベルの制約の有効性

提案手法では指定した語彙レベルより難しい単語を除外した発話生成用単語リストを作成することで，生成される発話の語彙レベルを制御しているが，生成された発話には指定した語彙レベルより難しい単語が含まれることがあった．各発話に対するそのような単語が含まれる割合のマクロ平均を表 2 に示す．“Base” の行はベースラインモデルの結果を表す．

LL が高いほど，難しい単語が多く使われる傾向があるが，これは難しい単語であるか否かは LL を基準にしているため，LL が高ければより多くの単語が難しいと判定されるからである．

$LL \geq 1.0$ の時，提案手法は $N = 25$ の時にベースラインモデルに比べて難しい単語の割合が低くなっているが，$N > 25$ の時はより多くの難しい単語が使われている．このことから，指定した語彙レベルが簡単な時に短い発話生成用単語リストは難しい単語の使用を減らす効果があると考えられる．

一方，$LL < 1.0$ の時，ベースラインモデルは提案手法より使われた難しい単語が少ないが，それはそもそも ChatGPT は難しい単語ではなく，よく出現する単語を使う傾向があるからだと考えられる．

表 2 生成された発話で使われた難しい単語の割合のマクロ平均

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
Base	1.6%	2.5%	4.2%	6.6%	8.9%
25	1.9%	2.8%	3.8%	5.8%	8.2%
50	1.9%	3.0%	4.1%	6.2%	8.9%
100	1.9%	3.0%	4.3%	6.5%	9.5%
200	1.8%	2.9%	4.3%	6.5%	9.4%
400	1.9%	2.8%	4.4%	6.7%	9.3%
800	1.9%	2.8%	4.5%	6.7%	9.3%
1600	1.8%	2.9%	4.3%	6.8%	9.5%

5.3 発話生成用単語の使用

表 3 生成された発話で使われた発話生成用単語の割合のマクロ平均

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
25	15.2%	16.8%	18.8%	21.3%	23.1%
50	12.4%	14.1%	16.7%	19.4%	22.2%
100	11.8%	13.9%	16.5%	20.0%	22.9%
200	12.5%	15.6%	19.1%	22.4%	23.3%
400	15.6%	18.6%	21.2%	22.8%	23.4%
800	20.0%	22.3%	22.7%	22.9%	23.5%
1600	22.5%	23.0%	22.8%	23.1%	23.4%

生成された発話における発話生成用単語リスト中の語の割合のマクロ平均を表 3 に示す．N が小さい時，LL が高ければより多くの発話生成用単語が使われるが，これは ChatGPT はよく出現する単語を使

う傾向があり，一定の発話生成用単語数 N に対して LL が高ければそのような単語がより多く含まれるからだと考えられる．

また，N を 25 から 100 に増やした時，使われた発話生成用単語は少なくなるが，これは ChatGPT が長い発話生成用単語リストを正しく処理できないためと考えられる．

N を更に増やすと，使われた発話生成用単語が再び増えるが，N が小さい時と異なり LL が高くなっても発話生成用単語の使用はあまり増えない．これは，N が大きい時は発話生成用単語リストには語彙レベルの制約に関わらず多くの単語が含まれているため，ChatGPT がリスト内の単語を偶然使う可能性が高くなるからだと考えられる．従って，ChatGPT に発話生成用単語を使わせるためには，短い発話生成用単語リストの方が有効であると言える．

5.4 生成された発話の語彙レベル

表 4 生成された発話で使われた最も難しい単語の語彙レベルのマクロ平均

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
(Base)	0.47	0.47	0.47	0.47	0.47
25	0.33	0.38	0.43	0.48	0.53
50	0.37	0.40	0.46	0.50	0.55
100	0.40	0.43	0.47	0.52	0.55
200	0.43	0.44	0.49	0.51	0.55
400	0.41	0.45	0.46	0.51	0.55
800	0.43	0.46	0.46	0.51	0.55
1600	0.41	0.43	0.46	0.50	0.55

表 5 生成された発話で使われた単語の語彙レベルの第一四分位数のマクロ平均

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
(Base)	1.05	1.05	1.05	1.05	1.05
25	1.00	1.04	1.09	1.14	1.17
50	1.03	1.05	1.09	1.13	1.16
100	1.04	1.06	1.09	1.12	1.15
200	1.05	1.06	1.09	1.12	1.15
400	1.05	1.06	1.08	1.11	1.15
800	1.05	1.07	1.08	1.12	1.15
1600	1.05	1.06	1.09	1.12	1.15

表 6 生成された発話で使われた最も簡単な単語の語彙レベルのマクロ平均

$N \setminus LL$	0.6	0.8	1.0	1.2	1.4
(Base)	1.77	1.77	1.77	1.77	1.77
25	1.82	1.82	1.82	1.82	1.82
50	1.81	1.81	1.81	1.81	1.81
100	1.80	1.80	1.80	1.80	1.80
200	1.79	1.80	1.79	1.80	1.80
400	1.80	1.80	1.80	1.80	1.80
800	1.80	1.80	1.80	1.80	1.80
1600	1.80	1.81	1.80	1.80	1.80

生成された発話で使われた最も難しい単語，難しさが上位 25%の単語，最も簡単な単語の語彙レベルのマクロ平均をそれぞれ表 4，表 5，表 6 に示す．使われた最も簡単な単語の語彙レベルは全ての (N, LL) ペアに対してほぼ同じであり，同じレベルの簡単な単語は常に使われると考えられる．一方，使われた最も難しい単語と難しさが上位 25%の単語の語彙レベルは LL に応じて変化したため，提案システムは生成された発話の語彙レベルに影響を与えたと言える．

また，語彙レベルは N が大きいほど簡単になり，発話生成用単語リストが長ければより簡単な単語がより多く使われることが分かる．

使われた最も難しい単語の語彙レベルが LL より低いことに関して，対話中に話者に一部分からない単語があっても問題なくコミュニケーションできるという我々の先行研究でも用いた仮定の元では問題にならない．

我々の先行研究では発話の語彙レベルは発話で使われた難しさが上位 25%の単語の語彙レベルとして定義したが，本論文で生成された発話に対して同じ定義で計算しても LL と一致しないことに関して，我々の先行研究では数百発話が考慮されるのに対し，ここでは一発話しか考慮されないため，単純には比較できない．そのため，生成された発話の語彙レベルを定量的に測定するために更なる評価指標が必要である．

5.5 結論

評価実験の結果から，提案手法で最適な発話生成用単語数 N は 25 であると結論付けられる．性能に関して，$N = 25$ の時 LL より難しい単語の使用が少なく (5.2 節)，発話生成用単語の使用が多く (5.3 節)，生成された発話の語彙レベルへの影響も多い (5.4 節)．また，コストに関して，発話生成用単語リストが短いためプロンプトが短く，発話の生成時間と費用も比較的少ない．

6 おわりに

本論文では，対話システムにおける語彙レベルを制御するためのモジュールを提案した．対話システムのベースモデルとして ChatGPT を採用し，雑談対話を用いて提案手法を評価し，発話生成用単語数 N と指定の語彙レベル LL による影響を考察した．

今後の課題として，生成された発話の語彙レベルを定量的に測定するための評価指標を提案し，語彙レベルをより精密に制御するために発話生成用単語リストの作成手法を改良する予定である．

謝辞

本研究は JSPS 科研費 JP21K18358 の助成を受けたものです.

参考文献

M. Asahara. Word familiarity rate estimation using a Bayesian linear mixed model. In S. Paun and D. Hovy, editors, **Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP**, pages 6–14, Hong Kong, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5902. URL https://aclanthology.org/D19-5902.

Y. Ehara. Innovative software to efficiently learn english through extensive reading and personalized vocabulary acquisition. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O. C. Santos, editors, **Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky**, pages 187–192, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-36336-8.

F. Landwehr, E. Varis Doggett, and R. M. Weber. Memories for virtual AI characters. In C. M. Keet, H.-Y. Lee, and S. Zarrieß, editors, **Proceedings of the 16th International Natural Language Generation Conference**, pages 237–252, Prague, Czechia, Sept. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.17. URL https://aclanthology.org/2023.inlg-main.17.

National Institute for Japanese Language and Linguistics. *Bunrui goihyo zôho kaitei-ban* (Word List by Semantic Principles Revised and Enlarged Edition). Dainippon Tosyo, Tokyo, 2004.

M. J. Pickering and S. Garrod. Alignment as the basis for successful communication. **Research on Language and Computation**, 4:203–238, 2006.

Y. Tseng, T. Tokunaga, and H. Yokono. Lexical level alignment in dialogue. In **Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers**, Maribor, Slovenia, Aug. 2023. SEMDIAL. URL http://semdial.org/anthology/Z23-Tseng_semdial_0010.pdf.

M. USAMI. Building of a Japanese 1000 person natural conversation corpus for pragmatic analyses and its multilateral studies, and ninjal institute-based projects: Multiple approaches to analyzing the communication of japanese language learners., 2023.

A　予備実験

ChatGPT が語彙レベルを理解しているか検証した．具体的には，二つの単語を ChatGPT に提示し，どちらがより簡単かを回答させた．

A.1　出題範囲

出題の候補は語彙分類表の分類コードの小数部分の同じ単語ペアとした．合計 519 分類のうち 6,378,126 ペアが対象となる．

語彙レベルの指標として「聞く」の単語親密度を採用し，親密度の高い方の単語が簡単であるとみなした．また，単語ペアの親密度の差が大きいほどどちらが簡単か判定しやすくなるため，どこまでの親密度の差を区別できるか分析するために単語ペアは親密度の差を 0.2 刻みでグループし，0.0 ~ 0.2 から 3.6 ~ 3.8 の 19 グループに分けた．各グループから親密度の差が第 5, 15, . . ., 95 百分位数となる単語ペアを抽出し，合計 187 個の単語ペアを出題対象とした．

A.2　プロンプト

問題のプロンプトはタスク指定，模範回答，問題本文の三つの部分から構成される (図 4).

(System) 対話中に日本語学習者にとって分かりやすい方の単語を答えてください。単語のラベル (”A”または”B”) のみ出力すること。文脈がなくても必ず”A”か”B”を選択すること。

(User) A: 仕事をする
　　B: 軼掌する
(Assistant) A

(User) A: 食べ物
　　B: 嘉肴

図 4　語彙レベル比較問題のプロンプト

問題本文において単語のラベルのバイアスを除去するために，各単語ペアに対して簡単な方の単語を A にする問題と難しい方の単語を A にする問題の二通りの問題を出題し，それぞれの正答率を分析した．

模範回答において正解の順番の影響を考慮し，0-shot の他に模範回答が ‘A”，“B”，“AB”，“BA”，“ABAB”，“BABA” の合計 7 種類の設定で実験を行った．

A.3　結果

ChatGPT の正答率を図 5 に示す．正答率が不安定であり，模範解答の提示順に回答が影響される傾向があるため，ChatGPT は語彙レベルの概念を上手く習得できていないと考えられる．

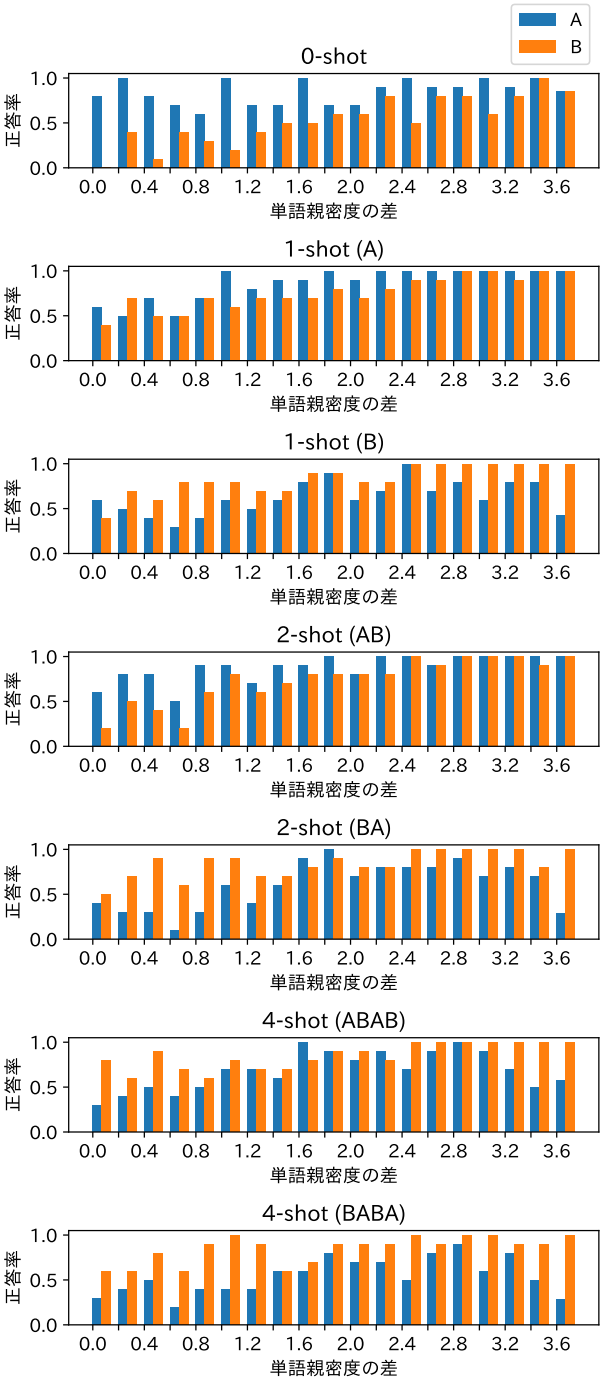

図 5　ChatGPT の語彙レベル比較問題の正答率．青色とオレンジ色の棒はそれぞれ正解が A と B の問題に対する正答率を表す．