言語特徴量導入によるソーシャルスキルレベル推定の性能向上

佐賀健志
奈良先端科学技術大学院大学
saga.takeshi.sn0@is.naist.jp

田中宏季
奈良先端科学技術大学院大学
hiroki-tan@is.naist.jp

岩坂英巳
奈良県立医科大学精神医学講座
iwasaka@heartland.or.jp

中村哲
奈良先端科学技術大学院大学
s-nakamura@is.naist.jp

1 はじめに

ソーシャルスキルは他者とのコミュニケーションにおいて重要な要素のひとつである。通常、私たちは幼少期にこれらのスキルを習得するが、自閉症スペクトラム症候群（ASD）や統合失調症などの影響で大人になってから困難に直面する人もいる [1]。ほとんどの日常生活はいくつかのソーシャルスキルの複合的な応用を必要とするため、そのようなスキルがなければ日常生活は非常に難しいものとなる。

この難しさを軽減する方法のひとつとして、ソーシャルスキルトレーニング（SST）がある。Bellack らによると SST は「条件付き反射療法」と「相互抑制による心理療法」、「社会的学習理論」に基づいており、様々な分野で広く使用されてきた [2, 3, 4, 5, 6]。また、SST は特定の疾患だけを対象としているわけではないため、精神疾患を抱える人だけではなく、ソーシャルスキルに悩みを抱える健常者でもソーシャルスキルレベル向上の効果が期待できる点も特徴のひとつである。

SST は一般的な手法になりつつあるが、効果的な SST を行えるようになるまでのトレーニングのハードルの高さなどの理由によりセラピストの数が不足しているため、受けたいと思った人がすぐには受けられないのが現状である [7]。この状況を改善するために、先行研究において SST 自動化が提案されてきた。

一般的な SST は「苦手スキルの明確化」「練習スキルの選択」「いい例と悪い例を見る」「ロールプレイ」「評価」「フィードバック」の順で行われる。SST の効果を最大化するためには「苦手スキルの明確化」が特に重要である。人対人の SST においてこの段階は口頭のコミュニケーションにおいて行われるが、現状の自動対話エージェントでの実現はまだまだ難しい。そのため著者らは先行研究において、マルチモーダル特徴量を入力として解釈性の高い線形回帰を用いてきた [8, 9]。本稿では既存手法に言語特徴量として BERT 埋込類似性（単語間・内容語間・文間）、Word2Vec 埋込類似性（単語間・内容語間・文間）、フィラー・代名詞・接続詞割合、隣接文間内容語一致割合を加え、特徴量選択を行うことでソーシャルスキルレベル推定性能の向上を目指す。

2 先行研究

先行研究において、誰でも SST を利用できるような環境実現を目指し、いくつかの仮想エージェントによる自動化が提案されている。

Tanaka らは仮想エージェントによるマルチモーダル対話システムを使用して、SST の自動化を試みた [10]。彼らの研究では、ユーザのスピーキングスキル向上を目的に設定し、スピーキングレベルを測定するための評価指標としてスピーキングスキルスコアを用いている。このスコアは経験豊富な数人のセラピストによってつけられた点数の平均値で、ユーザのソーシャルスキルレベルの推定モデルを学習させる際の正解ラベルとして用いられた。しかし、このスコアはセラピストの主観に基づいているため、個人のバイアスがかかっている可能性を含んでいる。

Voleti らは Social Skills Performance Assessment (SSPA) タスクにおいて撮影された動画から書き起こされたテキストのみを用いて、SSPA スコアを相関係数 0.752 で予測できることを示した [11]。しかし、使用したデータの SSPA スコアは 1 人のアノテータの主観に基づいているため、Tanaka らの場合と同様の問題を含む可能性が存在する。

これらの問題を解決するために本研究ではセラピストの主観に依存しない評価指標、対人応答性尺度に基

づいてモデルの学習・性能評価を行った。

3　対人応答性尺度 (SRS-2)

対人応答性尺度（SRS-2）は、65 項目の質問で構成される評価指標である。SRS-2 は元々、自閉症スペクトラム障害（ASD）を潜在的に抱えている人を評価するために設計されたが、統合失調症などの精神疾患を区別することもできることが報告されている [12]。また、健康な人々を対象とした実験においても被験者の ASD 傾向を判別できることが確認されているため、本尺度は信頼性の高い総合的評価指標として使用できると期待される。他者評定版と自己評定版が存在するが、本研究では自己評定版を使用した。

本研究では、SRS-2 の「総合スコア」とその治療下位尺度である「社会的コミュニケーション（65 項目のうち 22 項目）」を学習時の目的変数として利用した。SRS-2 の総合スコアには、純粋なコミュニケーションスキルだけではなく生活スタイルなどに関する項目も含まれるため、ユーザのソーシャルコミュニケーションスキルをより明確に示している社会的コミュニケーションスコアも使用することとした。

社会的コミュニケーションスコアと総合スコアは、相関係数 0.92 と高い相関であった。その一方で、先行研究で使用されていたセラピストの主観評価値であるスピーキングスコアとの相関係数は、それぞれ-0.19 と-0.29 という弱い負の相関であった [13]。また、SRS-2 は ASD 傾向を評価する目的で開発されたため、SRS スコアが高い人ほどソーシャルスキルレベルは低いことを意味する点に注意が必要である。

4　マルチモーダル特徴量

ソーシャルスキルは発言内容だけではなく表情や声の抑揚などの複合的な要素によって構成されているため、本研究では機械学習モデルへの入力としてマルチモーダル特徴量を使用する。基本的な特徴量は先行研究の特徴量セットをベースとし、言語特徴量を新しいものに入れ替えた。

4.1　音声・画像特徴量

先行研究と同様に、28 項目の音声特徴量と 26 項目の画像特徴量を使用している [9]。28 項目の音声特徴量は類似タスクである自動面接評価システムに使用されていたものを応用したものである [14][1)]。

1)　使用した各特徴量の詳細は付録を参照

表 1　言語特徴量

Feature name	Description
BERT_word	隣接単語間 BERT 埋込類似性
w2v_word	隣接単語間 Word2Vec 埋込類似性
BERT_sent	隣接文間 BERT 埋込類似性
w2v_sent	隣接文間 Word2Vec 埋込類似性
BERT_cont	内容語間 BERT 埋込類似性
w2v_cont	内容語間 Word2Vec 埋込類似性
Conj%	総単語数に対する接続詞割合
Filler%	総単語数に対するフィラー割合
Pronoun%	総単語数に対する代名詞割合
Neighbor_cont	隣接文間での内容語一致割合

4.2　言語特徴量

先行研究の入力特徴量において、音声特徴量・画像特徴量の数に比べて言語特徴量の数が少なかったため、予測に必要な言語情報を十分に抽出できていない可能性があった [9]。その問題を解決するため、アルツハイマー病の検出に使用されていた言語特徴量を参考に表 1 に示した特徴量を導入した [15]。元論文では Word2Vec 埋め込みに基づいた隣接発話間コサイン類似度の平均・最大・最小値を特徴量の一部として用いていたが、本研究では埋め込みを BERT 由来のものに、発話間類似度を文間・単語間・内容語間類似度に拡張している。

具体的な特徴量の算出方法は以下の通りである。実装にあたり BERT 埋込の算出には Transformers ライブラリの Whole-word-masking で事前学習された日本語 BERT モデルを使用し、Word2Vec 埋込算出には chiVe を使用している [16, 17, 18]。

4.2.1　BERT_word、w2v_word

まず、BERT と Word2Vec によって各単語に対する埋込ベクトル（単語レベル埋込）を算出する。その後、隣接単語間でコサイン類似度を計算する。それら類似度を発話テキスト全体の総単語数で平均することでこれらの特徴量は算出された。

4.2.2　BERT_cont、w2v_cont

まず、BERT と Word2Vec によって各内容語に対する埋込ベクトル（内容語レベル埋込）を算出する。事前学習済み BERT では、前処理で入力文が Subword に分割されるため、そのままでは内容語と BERT 埋込ベクトルとの対応が取れない。そのため、形態素解析

器 Sudachi で分かち書きされた単語と BERT-tokenizer で分割された sub-word で文字レベル一致度を算出し、内容語に対して 0.25 を超えたものを内容語 BERT 埋込ベクトルとして使用した [19]。そして、隣接内容語間でコサイン類似度を計算し、それらを発話テキスト全体の総内容語数で平均することでこれらの特徴量は算出された。

4.2.3 BERT_sent、w2v_sent

単語レベル埋込の隣接単語間コサイン類似度を算出後、各文に含まれる総単語数で平均をとる（文レベル類似度）。それら文レベル類似度を発話テキスト全体の総文数で平均することでこれらの特徴量は算出された。

4.2.4 Conj%、Filler%、Pronoun%

まず、Sudachi によって文を単語に分割した後、原型に戻す。その後、発話テキスト全体における総単語数に対する接続詞・フィラー・代名詞の割合を算出することでこれら特徴量は得られた。

4.2.5 Neighbor_cont

まず、Sudachi によって文を単語に分割した後、原型に戻す。その後、隣接文間での内容語の一致割合を算出することでこれら特徴量は算出された。

5 実験

5.1 学習データ

学習データには、先行研究によって得られた日本語の 1 分間スピーキングビデオを使用した [13]。まず、被験者はノートパソコンの画面を介して MMDAgent（http://www.mmdagent.jp/）で作成された仮想エージェントに対して、最近あったうれしかった出来事について 1 分間話す。被験者が話している間、被験者の顔の映像と声が記録され、後日、アノテータによってテキストに書き起こされた。1 分間の発話終了後、SRS 質問紙を用いて、被験者の生活スタイルやコミュニケーションスタイルなどについて記入してもらった。以上のような手順によって 27 人の健康な被験者のデータセットは作成された。うち 1 名のデータにて特徴量抽出における外れ値が確認されたため、それを除外した 26 名のデータを今回は使用している。

表 2　SRS スコアと特徴量との相間係数

Name	SRS	Com
BERT_word	0.16	0.18
w2v_word	-0.20	-0.12
BERT_sent	-0.14	-0.03
w2v_sent	0.03	0.08
Conj%	0.10	0.09
Filler%	0.22	0.21
Pronoun%	-0.08	-0.08
BERT_cont	0.19	0.31
w2v_content	-0.07	-0.11
Neighbor_cont	-0.03	-0.02

5.2 相関分析

まず、各特徴量と SRS との間の関連性を調べるために相関分析を行った。その結果を表 2 に示す。ここで SRS は総合スコア、Com は社会的コミュニケーションスコアを示している。いずれの特徴量においても、SRS スコアとの有意な相関は見られていない ($p>0.05$)。

単語レベルの BERT_word や内容語レベルの BERT_cont、Filler%で比較的高い正の相関が確認できる。その一方で単語レベルの w2v_word は比較的強い負の相関を示しており、BERT を用いた特徴量とは異なる言語的特徴を取得している可能性が示唆される結果となった。

5.3 特徴量選択

一般に線形回帰のような単純な機械学習モデルに対して特徴量選択を行うことで推定精度が向上することが知られている。本研究でもそれに倣い、各特徴量と SRS スコアの Pearson 相関係数に基づいた Filter Method によって特徴量選択を行った後、線形回帰で SRS スコア予測を試みた。先行研究において 10 個の特徴量で線形回帰を行っていたため、それに倣い本研究では上位 10 個の特徴量選択を行うこととした [10]。提案手法に対する特徴量選択によって選ばれた特徴量を表 4 に示す。また、特徴量選択を行う場合と行わなかった場合の特徴量セットを用いた線形回帰による予測結果を表 3 に示す。ここで RMSE は真値と予測値のずれを表す二乗平均平方根誤差、Correl は Pearson 相関係数を示している。比較のために従来手法に対して同様の操作を行った場合の結果も示している [9]。

表 3　SRS スコア予測結果

	SRS			Com		
Method	RMSE	Correl	p-value	RMSE	Correl	p-value
先行手法 [9]	18.66	0.45	0.02	8.19	0.56	0.003
先行手法 [9] ＋特徴量選択	15.03	0.72	2.60e-5	8.15	0.66	2.55e-4
提案手法	20.44	0.28	0.17	10.08	0.29	0.15
提案手法＋特徴量選択	13.49	0.76	6.82e-6	10.53	0.44	0.024

表 4　特徴量選択結果（上位 10 個）

Rank	SRS	Com
1	F0_SD	F0_SD
2	AU17	AU45
3	AU45	F3_SD
4	F3_SD	AU17
5	Pose_Rx_CV	AU28
6	Pose_Rx	AU23
7	AU28	Pose_Rx_CV
8	AU15	F3_freq_mean
9	AU01	BERT_cont
10	Pose_Rz_CV	AU15

特徴量選択をしなかった場合、いずれの指標においても性能が低下する結果となった。その一方で特徴量選択を行った場合、総合スコア予測において性能が向上し、今回試した全ての手法の中で最も高い相関と低い RMSE を達成した。参考までに特徴量選択後の提案手法で SRS 総合スコアを予測した場合の真値と予測値の散布図を図 1 に示す。

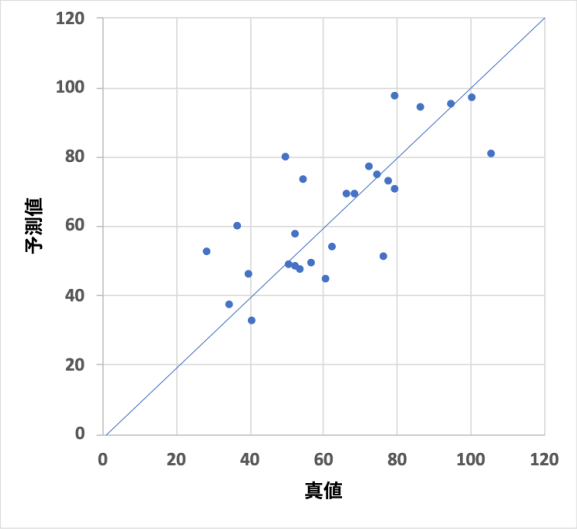

図 1　提案手法による総合スコア予測（r : 0.76）

6　考察

相関分析において Filler%が比較的高い正の相関を示しており、ソーシャルスキルと関連している特徴量である可能性が示された。先行研究のスピーキングスコア予測でもフィラーは重要特徴量として挙げられており、今回の結果によって SRS スコア予測でも同様のことが言えることがわかった [10]。

表 3 において相関係数が向上していることから、特徴量選択が推定性能の向上に大きく寄与していることが確認できる。特に、SRS 総合スコア予測においては提案手法に対して特徴量選択を行った場合が最も高い相関係数となっており、有効性が示唆されている。一方で、社会的コミュニケーションスコア予測に関しては、提案手法に特徴量選択をおこなった場合よりも既存手法に対して特徴量選択を行った場合のほうが有効であることが確認された。

表 4 に示されている上位 10 個の選択された特徴量を比較すると、社会的コミュニケーションスコア予測を行った場合にのみ、言語特徴量の BERT_cont が現れていることが確認できる。このことから全体のスコア予測には言語的内容が重要ではない可能性が示された。

7　おわりに

本研究では既存手法に新たな特徴量を追加し、特徴量選択を行うことで SRS スコア推定性能向上を実現した。また、特徴量選択の結果、ソーシャルスキルレベル推定において重要特徴量となりうる候補が挙げられた。今後は実例と照らし合わせながらより詳細な分析を行う必要がある。

8　謝辞

本研究は、CREST 戦略的創造研究推進事業 (JPMJCR19A5) および JSPS 科研費 (JP17H06101 および JP18K11437) の支援を受けて行われたものである。

参考文献

[1]Children and Adults with Attention-Deficit/Hyperactivity Disorder. Relationships & social skills. Accessed August 15, 2020.

[2]A. Salter. *Conditioned reflex therapy*. Creative Age Press, 1949.

[3]J. Wolpe. *Psychotherapy by reciprocal inhibition*. Stanford University Press, 1958.

[4]A. Bandura. *Principles of behavior modification*. Holt, Rinehart and Winston, 1969.

[5]Kim T. Mueser and Alan S. Bellack. Social skills training: Alive and well? *Journal of Mental Health*, pp. 549–552, 2007.

[6]A. S. Bellack, K. T. Mueser, S. Gingerich, and J. Agresta. *Social Skills Training for Schizophrenia: A Step-by-Step Guide*. Guilford Press, 2 edition, 2004.

[7]一般社団法人 SST 普及協会. 認定講師. Accessed October 5, 2020.

[8]T. Saga, H. Tanaka, H. Iwasaka, and S. Nakamura. Objective prediction of social skills level for automated social skills training using audio and text information. In *ICMI '20 Companion: Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 467–471. Association for Computing Machinery, New York, NY, United States, 2020.

[9]佐賀健志, 田中宏季, 岩坂英巳, 中村哲. マルチモーダル情報を用いたソーシャルスキルの客観的推定. HCG シンポジウム 2020, pp. A–7–2. 電子情報通信学会, 2020.

[10]Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE*, Vol. 12, No. 8, pp. 1–15, 08 2017.

[11]R.Voleti, S.Woolridge, J.M.Liss, M.Milanovic, C.R.Bowie, and V.Berisha. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. In *Proceedings of INTERSPEECH2019*, pp. 1433–1437, 2019.

[12]MD John N. Constantino and PhD Christian P. Gruber. *Social Responsiveness Scale, Second Edition (SRS-2)Back*. Western Psychological Services, 2012.

[13]H. Tanaka, H. Iwasaka, H. Negoro, and S. Nakamura. Analysis of conversational listening skills toward agent-based social skills training. *Journal on Multimodal User Interfaces*, Vol. 14, No. 1, pp. 73–82, Mar 2020.

[14]I. Naim, M. I. Tanveer, D. Gildea, and M. E. Hoque. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, Vol. 9, No. 2, pp. 191–204, April 2018.

[15]A.Balagopalan, B.Eyre, F.Rudzicz, and J.Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimers disease detection. In *Proceedings of INTERSPEECH2020*, pp. 2167–2171, 2020.

[16]Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing, 2019.

[17]真鍋陽俊, 岡照晃, 海川祥毅, 髙岡一馬, 内田佳孝, 浅原正幸. 複数粒度の分割結果に基づく日本語単語分散表現. 言語処理学会第 25 回年次大会 (NLP2019), pp. NLP2019–P8–5. 言語処理学会, 2019.

[18]河村宗一郎, 久本空海, 真鍋陽俊, 髙岡一馬, 内田佳孝, 岡照晃, 浅原正幸. chive 2.0: Sudachi と nwjc を用いた実用的な日本語単語ベクトルの実現へ向けて. 言語処理学会第 26 回年次大会 (NLP2020), pp. NLP2020–P6–16. 言語処理学会, 2020.

[19]K. Takaoka, S. Hisamoto, N. Kawahara, M. Sakamoto, Y. Uchida, and Y. Matsumoto. Sudachi: a japanese tokenizer for business. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).

A　付録：使用した特徴量一覧

表 A.1 に使用した特徴量一覧を示す。

表 **A.1**　使用した特徴量（上から音声・画像）

Feature name	Description
Energy	平均スペクトルエネルギー
F0, F1, F2, F3 Mean	F0,F1,F2,F3 の平均周波数
F0, F1, F2, F3 SD	F0,F1,F2,F3 の標準偏差
F0 Min, Max	F0 周波数の最小・最大値
F0 range	F0 周波数の最大最小値の差
F1, F2, F3 BW	F1,F2,F3 の平均バンド幅
F2/F1, F3/F1 Mean	F2-F1, F3-F1 の平均比率
F2/F1, F3/F1 SD	F2-F1 比, F3-F1 比の標準偏差
Int mean,SD	音声強度の平均・標準偏差
Int Min, Max	最小・最大音声強度
Int range	最小最大音声強度の差
Jitter, Shimmer	F0 周波数・音声強度のゆらぎ
Unvoiced, Breaks %	無声区間・短時間無声区間の比率
AU01, AU02	眉の内側・外側を上げる
AU04	眉を下げる
AU05	上瞼を上げる
AU06	頬を持ち上げる
AU07	瞼を緊張させる
AU09	鼻にしわを寄せる
AU10	上唇を上げる
AU12	鼻唇溝を深める
AU14	えくぼを作る
AU15	唇両端を下げる
AU17	オトガイを上げる
AU20	唇両端を横に引く
AU23	唇を固く閉じる
AU25	顎を下げずに唇を開く
AU26	顎を下げて唇を開く
AU28	唇を吸い込む
AU45	まばたく
AU06+12	幸せな感情と相関する AU の組み合わせ
Pitch,Yaw,Roll	頭の回転角度