ユーザ情報抽出のための自己開示文の人物属性分類

平野 徹　小林 のぞみ　東中 竜一郎　牧野 俊朗　松尾 義博
日本電信電話株式会社 NTT メディアインテリジェンス研究所
{hirano.tohru, kobayashi.nozomi, higashinaka.ryuichiro, makino.toshiro, matsuo.yoshihiro}@lab.ntt.co.jp

1 はじめに

近年，エンターテイメントやカウンセリングなどを目的とした，雑談を行う対話システムの研究が増加している [3, 11]．雑談対話では，話者の好みや経験といった，話者自身に関する情報が述べられることがある．我々の調査では，人間同士の雑談対話における約 26%の発話が話者自身に関する情報を述べた自己開示文であることが分かった．またこの傾向は対話システムが対話相手である場合も同様であった．

我々は，雑談対話で述べられる話者自身に関する情報を活用することにより，雑談対話のパーソナライズだけでなく，映画館案内等のタスク対話や新商品等の情報提供においてもパーソナライズ可能な対話エージェントの実現を目指している．例えば，過去に「イギリスに行ってきたよ」「横須賀に住んでるよ」「西野カナが好きだよ」という発話をしたユーザに対して，以下のようなパーソナライズされた対話を行うエージェントを実現したい．

雑談対話 前の会話を覚えていることを伝える
USER：「旅行に行きたいな」
SYSTEM：「この間，**イギリスに行って**たね」

タスク対話 未指定条件をユーザ情報で補う
USER：「アナ雪の上映時間は？」
SYSTEM：「**横須賀映画館**では 10 時からです」

情報提供 ユーザが興味ある情報を提供する
USER：「そろそろカラオケの練習しないと」
SYSTEM：「**西野カナ**の新曲が来週発売されるよ」

これらのパーソナライズされた対話を行うエージェントを実現するためには，ユーザ発話から次の要件を満たすユーザ情報を抽出する必要があると考えた．

1. 雑談対話におけるシステム発話を生成するための情報があること
2. タスク対話の未指定条件を補うための情報があること
3. 情報提供する情報を検索するための情報があること
4. どのユーザ情報をいつ利用するか判断するための情報があること

本研究では，要件 1 を満たすために，「何がどうした」を示す**述語項構造**を抽出する．述語項構造からシステム発話を生成する方法については文献 [3] で提案されている．次に，要件 2 と要件 3 を満たすために，キーワードとなる人名や地名等の **Entity** を抽出する．最後に，要件 4 のどの情報を利用するか判断する情報として，経験や居住地，趣味といった**人物属性**を，いつ利用するか判断する情報として，旅行や住まい，音楽といった**トピック**を抽出する．なお，先の例では，直前発話と同じトピックのユーザ情報が用いられている．

以上のことから，雑談対話におけるユーザの発話文から，ユーザ自身に関する情報を〈**述語項構造，Entity，人物属性，トピック**〉の構造化された形で抽出することを試みる．上述の「イギリスに行ってきたよ」の発話からは〈(行く ニ:イギリス), イギリス, 経験, 旅行〉を，「横須賀に住んでるよ」の発話からは〈(住む ニ:横須賀), 横須賀, 居住地, 住まい〉を，「西野カナが好きだよ」の発話からは〈(好き ガ:西野カナ), 西野カナ, 趣味, 音楽〉をそれぞれ抽出する．

ユーザ情報抽出の処理は，まず，ユーザ発話がユーザ自身に関する情報が述べられている自己開示文か判定する．次に，自己開示文と判定されたユーザ発話から，述語項構造，Entity，人物属性，トピックを抽出する．自己開示文の判定には対話行為推定 [9] が，述語項構造の抽出には文献 [6] の述語項構造解析が，Entity の抽出には焦点語抽出 [3] が，トピックの抽出にはトピック分類 [12] が適用可能であると考えられるため，本稿では最後の課題である人物属性の抽出について取り組む．

自己開示文から人物属性を抽出するタスクを，既定の人物属性カテゴリから最適なカテゴリを 1 つ選択する分類問題とする．分類問題においては，トピック分類のように主に内容語を用いた分類手法が利用されているが，人物属性分類においては，内容語だけでは区別できないカテゴリがある．そこで本稿では，この課題を解決するために，従来の内容語の手掛りに加え，機能語と副詞の意味情報を手掛りにした分類手法を提案し，その有効性

について議論する．

2 タスク設定

対話におけるユーザの自己開示文には，これまでに示した例のように，当該自己開示文のみで人物属性が判断できるものと，当該自己開示文の直前文も読まないと判断できないものがある．後者は，以下のように，直前の質問文への回答や直前の自己開示文への対比などに多い．

質問文への回答

SYSTEM：「好きな歌手は誰?」
USER：「西野カナ」

自己開示文への対比

SYSTEM：「ミスチルが好きです」
USER：「私は西野カナ」

本研究では，このような直前文を読まないと判断できない自己開示文も正しく人物属性分類を行うために，自己開示文とその直前文の 2 文を入力とする．

次に，人物属性カテゴリについて述べる．人物属性カテゴリの参考となる研究として，大規模に人物属性に関する質問文を収集し分析した研究 [14] がある．この文献では，収集した質問文を,「名前」「出身地」等の約 1000 個の質問文カテゴリとしてまとめている．

この質問文カテゴリには「野球が好き」「サッカーが好き」のように Entity が含まれたものが多数存在するが，全ての Entity を網羅できる体系にはなっていない．質問文カテゴリを人物属性カテゴリとみなし，人手で雑談対話中の自己開示文へカテゴリを付与すると，カテゴリを付与できた自己開示文は 32%のみであった．

そこで，一般的な市場調査アンケートを参考にしつつ，質問文カテゴリから Entity を除き，表 1 に示す 34 種類の人物属性カテゴリを設定した．本カテゴリを再度雑談対話中の自己開示文 (18,579 文) へ付与した結果，全自己開示文にカテゴリが付与できた．表 1 に各カテゴリが付与された頻度を示す．

また作業者間の一致率を調べるために，ランダムに抽出した自己開示 200 文へ別の作業者がカテゴリ付与作業を行った．作業者間の判断が一致したのは 181 文 (κ 値 =0.885) と，一致率が 90.5%と高いことがわかった．

上記の結果より，ユーザの自己開示文とその直前文の 2 文を入力とし，34 種類の人物属性カテゴリから最適なカテゴリを 1 つ選択する問題は，人が共通の指針を持って判断できることを確認でき，妥当なタスク設定であると考えられる．

3 提案手法

上述した人物属性分類タスクにおいて，トピック分類のような主に内容語を用いた分類手法では,「さっきテニスをした」という経験と「いつもテニスをしている」という習慣などのカテゴリを誤って分類することがあり，分類精度低下の要因になっている．

表 1: 人物属性カテゴリと雑談対話における頻度

	人物属性カテゴリ	頻度
1	人間関係 (家族 配偶者以外)	55
2	人間関係 (配偶者・結婚)	40
3	人間関係 (恋人・恋愛)	28
4	人間関係 (他人 第三者)	39
5	名前	25
6	性別	17
7	年齢	27
8	血液型	22
9	誕生日	18
10	星座	12
11	干支	12
12	性格	98
13	身体的特徴	55
14	出身地	43
15	居住地	96
16	同居者	53
17	住居区別	26
18	職業	104
19	勤務地	13
20	役職	13
21	通勤通学	36
22	経歴	91
23	収入	22
24	支出	81
25	所有物	343
26	知識	491
27	趣味・嗜好	2,234
28	習慣・行動特性	2,414
29	経験・記憶	4,758
30	特技・得意	157
31	能力	208
32	意見・感想	5,580
33	願望	764
34	その他	604
	合計	18,579

表 1 で示した人物属性カテゴリにおいて，人間関係から知識までの 26 種類のカテゴリの分類には内容語は重要な手掛りになるが，趣味以降のカテゴリの分類には手掛りとならない．例えば，先の例はどちらも同じ内容語「テニス」「する」であることからも内容語からは判別し難いことがわかる．これらの人物属性カテゴリ，特にユーザの行動に関するカテゴリを正しく分類するためには，行動が完了しているか，継続しているか，繰り返されているかなどの情報をとらえる必要がある．そこで，従来の内容語の手掛りに加え，機能語や副詞の意味情報を手掛りにした分類手法を提案する．

3.1 機能語の意味情報

提案手法では,「た」「ている」等の機能語の意味情報を利用する．例えば,「た」は完了,「ている」は継続の意味を持っているので，これらを分類の手掛かりとする

ことで経験，習慣を区別可能となる．本稿では，今村ら [5] の手法を用いて機能語の意味を解析し得られた意味ラベルを手掛りとして利用する．特に，人物属性分類においては,「継続」が習慣・行動特性,「完了」が経験・記憶,「可能」が能力,「推量」「感嘆」が意見・感想,「願望」「依頼」が願望の分類に寄与すると期待できる．

ユーザの自己開示文およびその直前文から抽出した機能語の意味ラベルは，次の 2 つの方法で人物属性分類の手掛りとして利用する．1 つ目に，抽出された意味ラベルを，自己開示文と直前文のどちらから抽出されたものかを区別して素性化する．具体的には,「さっきテニスをした」の自己開示文から完了の意味ラベルが抽出された場合は，“自己開示文_完了” のように，“⟨ 抽出文 ⟩_⟨ 意味ラベル ⟩” という素性とする．

2 つ目に，意味ラベルが抽出された機能語の出現位置に基づき，意味ラベルに順位を付けて素性化する．これは，1 つの文から複数の意味ラベルが抽出された場合，文末に近い意味ラベルのほうが人物属性分類の重要な手掛りになること，自己開示文に意味ラベルが存在しない場合は，直前文の意味ラベルが人物属性分類の重要な手掛りとなることに鑑みた方法である．具体的には，直前文から「可能」が，自己開示文から「継続」「完了」の順で意味ラベルが抽出された場合，文末に近い意味ラベルから順に，“1 位_完了”，“2 位_継続”，“3 位_可能” のように，“⟨ 順位 ⟩_⟨ 意味ラベル ⟩” という素性とする．

3.2 副詞の意味情報

提案手法では，述部を修飾する副詞の情報として,「さっき」「いつも」等の副詞が持つ意味情報を利用する．例えば,「いつも」は修飾される行動が日常的に繰り返し行われていることを意味し,「さっき」は修飾される行動を行った時間を特定する意味を持っている．人物属性分類においては，日常的な繰り返しを意味する副詞が習慣・行動特性，時間を特定する意味を持つ副詞が経験・記憶の分類に寄与すると期待できる．

副詞の意味情報抽出には，あらかじめ用意した 2 種類のリスト，A) 日常的な繰り返しを意味する副詞のリスト (いつも，よく，等) と，B) 時間を特定する意味を持つ副詞のリスト (さっき，まだ，等) を利用する．抽出された副詞のタイプを，自己開示文と直前文のどちらから抽出されたものかを区別して素性化する．具体的には,「いつもテニスをしている」の自己開示文から副詞タイプ A が抽出された場合は，“自己開示文_副詞タイプ A” のように，“⟨ 抽出文 ⟩_⟨ 副詞タイプ ⟩” という素性とする．

表 2: 人物属性分類の正解率

	正解率
従来手法	76.0% (14,120/18,579)
提案手法	**88.9**% (16,523/18,579)
上限値 (参考)	90.5% (181/200)

4 評価実験

従来の主に内容語を手掛りとした手法と，機能語や副詞の意味情報も手掛りとした手法を比較し，提案手法の有効性を検証する．比較対象となる従来手法は，入力文中の全ての単語表記と日本語語彙大系 [4] のカテゴリを手掛りとした．なお，いずれの手法も LIBLINEAR[1] を利用し，ロジスティック回帰で学習した．

4.1 実験データ

本実験では，人と人がテキストチャット形式で雑談した 3,680 対話（71,422 発話）と，人とシステム [3] がテキストチャット形式で雑談した 480 対話（7,444 発話）の合計 4,160 対話中の自己開示文 18,579 文に対して，2 節で述べた人物属性カテゴリを人手で付与したデータを用いる．各人物属性カテゴリが付与された自己開示文数を表 1 に示す．最も多く付与されたカテゴリは，意見・感想で 5,580 回 (30%)，2 番目は経験・記憶で 4,758 回 (25%)，3 番目は習慣・行動特性で 2,414 回 (12%)，4 番目は趣味・嗜好で 2,234 回 (12%) となった．なお実験は 10 分割交差検定で実施し，人手で付与した人物属性カテゴリを正しく選択できるか評価した．

4.2 実験結果

実験結果を表 2 に示す．なお正解率とは，システムの分類結果がどの程度正しいかを示す尺度であり，次式の通りである．

$$\text{正解率} = \frac{\text{システムが正しく分類できた自己開示文数}}{\text{自己開示文数}}$$

実験結果から，提案手法は 88.9%と，従来手法の 76.0%に比べて，12.9 ポイント向上したことがわかった．また McNemar 検定を行い，提案手法の有効性を確認した．

提案手法によって正しく分類できるようになった事例を分析すると，習慣・行動特性，経験・記憶のカテゴリにおける改善が多く見られた．上記カテゴリの個別の正解率を見てみると,「継続」「完了」の意味ラベルを持つ機能語の手掛りと,「繰り返し」「時間特定」の意味を持つ副詞の手掛りによって，習慣・行動特性では 66.4%から 86.9%へ，経験・記憶では 68.9%から 89.0%へ向上した．

また，2 節で述べたとおり，本タスクの人手の一致率は 90.5%であり，これを上限値とすると，提案手法は上限値とほぼ同じ正解率になっていることが確認できる．

[1] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

5 関連研究

ユーザ発話からユーザ自身に関する情報を抽出する研究として [2, 8, 15] 等がある．文献 [15] は，雑談対話のパーソナライズのため，ALICE[2] と同様に，ユーザ発話から名前や趣味等のユーザ情報を抽出するルールを用意し，ルールにマッチした文字列を Entity として抽出する手法を用いている．抽出された Entity とテンプレートを利用して，“You like apple, John.” といったシステム発話を生成しているが，抽出されるユーザ情報が限定的であるという課題がある．

情報提供のパーソナライズのための研究 [2] は，同僚についての情報を提供する対話システムにおいて，ユーザ発話からユーザの座席の位置情報を抽出し，その位置情報を用いて DB 検索を行うことでパーソナライズされた情報提供を実現している．しかし，この研究も抽出されるユーザ情報が限定的であるという課題がある．

この課題を解決するために，文献 [8] は，ユーザ発話から ⟨I, like, apple⟩ のように 2 つの名詞句 (I, apple) とその間の意味的関係 (like) を示す表現からなる 3 つ組を抽出する手法が用いている．抽出された 3 つ組に基づいて，“I know, you like apple.” といったシステム発話を生成している．この研究では，雑談対話のパーソナライズに主眼をおいているため，トピックや人物属性は抽出していない．我々の研究は，パーソナライズの適用先をタスク対話や情報提供に拡張するための研究と位置づけることができる．

パーソナライズの研究は，情報検索 [10, 13] やレコメンド [1, 7] の分野で多く行われている．これらの研究では，ユーザの興味を単語ベクトル等で表現し，情報検索やレコメンドの対象となる情報の単語ベクトルとユーザの興味ベクトルとを比較して，ユーザの興味ベクトルに近い情報を提示することでパーソナライズを実現している．これらの手法では，ユーザの興味を大まかに捉えることはできるが，本研究のようにユーザの細かな情報を捉えることはできない．

6 おわりに

本稿では，ユーザ発話からユーザ自身に関する情報を ⟨ 述語項構造, Entity, 人物属性, トピック ⟩ の構造化された形で抽出することを目指し，自己開示文の人物属性分類タスクに取り組み，従来手法のように，内容語を中心とした手掛りだけでなく，機能語と副詞の意味情報に基づく手掛りも用いた分類手法を提案した．評価実験では，提案手法は従来手法よりも高い正解率で人物属性を分類できることを確認できた．

今後は，抽出したユーザ情報に基づいてパーソナライズ可能な対話システムを実装し，対話システムとしての評価を実施したい．またユーザ情報抽出の更なる精度向上にも取り組む予定である．

[2] http://www.alicebot.org/

参考文献

[1] Ardissono, L., Gena, C., Torasso, P., Bellifemine, F., Difino, A. and Negro, B.: User Modeling and Recommendation Techniques for Personalized Electronic Program Guides, *Personalized Digital Television - Targeting Programs to Individual Viewers*, Human - Computer Interaction Series, Vol. 6, pp. 3–26 (2004).

[2] Corbin, C., Morbini, F. and Traum, D.: Creating a Virtual Neighbor, *Proceedings of the 2015 International Workshop Series on Spoken Dialogue Systems Technology* (2015).

[3] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T. and Matsuo, Y.: Towards an Open Domain Conversational System Fully Based on Natural Language Processing, *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 928–939 (2014).

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: 日本語語彙大系 CD-ROM 版, 岩波書店 (1999).

[5] 今村賢治, 泉朋子, 菊井玄一郎, 佐藤理史: 述部機能表現の意味ラベルタガー, 言語処理学会第 17 回年次大会論文集, pp. 308–311 (2011).

[6] 今村賢治, 東中竜一郎, 泉朋子: ゼロ代名詞照応付き述語項構造解析の対話への適応, 言語処理学会第 20 回年次大会論文集, pp. 709–712 (2014).

[7] Jiang, Y., Liu, J., Tang, M. and Liu, X.: An Effective Web Service Recommendation Method Based on Personalized Collaborative Filtering, *Proceedings of the 2011 IEEE International Conference on Web Services*, pp. 211–218 (2011).

[8] Kim, Y., Bang, J., Choi, J., Ryu, S., Koo, S. and Lee, G. G.: Acquisition and Use of Long-term Memory for Personalized Dialog Systems, *Proceedings of the 2014 Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction* (2014).

[9] 目黒豊美, 東中竜一郎, 杉山弘晃, 南泰浩: 意味属性パターンを用いたマイクロブログ中の発言に対する自動対話行為付与, 情報処理学会研究報告音声言語情報処理, pp. 1–6 (2013).

[10] Qiu, F. and Cho, J.: Automatic Identification of User Interest for Personalized Search, *Proceedings of the 15th International Conference on World Wide Web*, pp. 727–736 (2006).

[11] Ritter, A., Cherry, C. and Dolan, W. B.: Data-Driven Response Generation in Social Media, *Proceedings of the Conference on Empirical Methods in Natural Language*, pp. 583–593 (2011).

[12] Sebastiani, F.: Text Categorization, *Text Mining and its Applications* (Zanasi, A.(ed.)), WIT Press, pp. 109–129 (2005).

[13] Shen, X., Tan, B. and Zhai, C.: Implicit User Modeling for Personalized Search, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 824–831 (2005).

[14] 杉山弘晃, 目黒豊美, 東中竜一郎, 南泰浩: 対話システムのパーソナリティを問う質問に対する発話生成, 人工知能学会言語・音声理解と対話処理研究会, pp. 33–38 (2014).

[15] Weizenbaum, J.: ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine, *Communications of the Association for Computing Machinery*, Vol. 9, pp. 36–45 (1966).