介護用対話システムのための高齢者の発話理解

淺尾 仁彦[1]　Julien Kloetzer[1]　水野 淳太[1]　齊木 大[2]　門脇 一真[1,2]　鳥澤 健太郎[1,3]

[1] 国立研究開発法人情報通信研究機構　[2] 株式会社日本総合研究所　[3] 奈良先端科学技術大学院大学

[1]{asao, julien, junta-m, kadowaki, torisawa}@nict.go.jp　[2]saiki.dai@jri.co.jp

1　はじめに

近年、少子高齢化に伴い、介護人材の不足が懸念されている。現在、内閣府総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム (SIP) の支援のもと、KDDI 株式会社・情報通信研究機構・NEC ソリューションイノベータ株式会社・株式会社日本総合研究所の共同で、介護に関わる人の負担を軽減するために、現在、高齢者の健康状態をモニタリングしたり、話し相手になったりすることのできる対話システムを構築するプロジェクトを進めている。本研究では、このプロジェクトの一部として、高齢者の発話を理解するための (i) 介護質問応答分類モジュールと (ii) 介護情報含意認識モジュールという 2 つのモジュールを BERT [1] を用いて構築し、その評価を行った。

介護質問応答分類モジュールは、あらかじめ用意した「悪寒はありますか」「週に 1 回以上外出していますか」など、高齢者の健康状態をモニタリングする質問に対して、高齢者の応答が「はい」を意味するのか「いいえ」を意味するのか、あるいはどちらでもないのかなどを分類するモジュールである。高齢者が間接的な答え方をした場合にも、意図を正しく認識することが目標である（例えば「週に 1 回以上外出していますか」に対して「毎週水曜に医者まで行ってますよ」と答えた場合に、これが「はい」を意味していることを認識できる必要がある）。一方、介護情報含意認識モジュールは、システムが質問をしていなくても、ユーザが自発的に介護情報を提供したときに、それを認識するためのモジュールである。例えば「食欲はありますか」という質問に対してユーザが「もりもり食べてますよ。最近調子いいです。夜もぐっすり寝てますし」と答えた場合、ユーザは食欲に関する情報だけでなく、まだ質問していない睡眠に関する情報も提供している。このような場合に、介護情報含意認識モジュールは「夜もぐっすり寝てますし」という発話が〈不眠の症状はない〉という介護情報を含意している（従って、「夜はよく眠れていますか」という質問はする必要がない）ことを認識することができる。2 つのモジュールの使用イメージを図 1 に示した。

上記 2 つのモジュールの実装に機械学習を利用するため、質問応答および自発的発話の学習データを作成した。学習データの作成にあたっては、BERT に基づくトピックワードモデルを利用して含意関係を効率的に発見する手法を提案する。

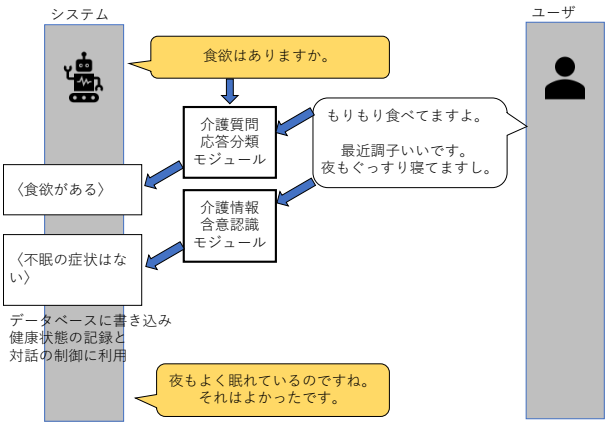

図 1: 本研究で開発したモジュールの利用イメージ

作成したデータに基づいて BERT ベースのニューラルネットワークモデルを学習し評価を行った。介護質問応答分類モジュールについては、データ作成において「はい」「そうです」などの単純な応答を避け、意図的に難しいデータを作成したにもかかわらず、平均精度 89.0%を達成した。介護情報含意認識モジュールは平均精度 87.1%を達成した。本研究で開発したモジュールで適切に処理できる対話の事例を表 1 に示す。

表 1: 開発したモジュールで適切に扱える対話の事例

介護質問応答分類モジュール
システム：お医者さんや薬局の人から薬を飲むときの注意点って説明してもらいましたか **ユーザ**：必ずお水か白湯で飲んでくださいねって **モジュールの判定**：YES
システム：お住いの近辺で、何かの活動をしたり何かに参加することってありますか **ユーザ**：あまり興味がわかないんですけど **モジュールの判定**：NO
介護情報含意認識モジュール
ユーザ：うがいなんて、気持ち悪い時にしかしませんよ **モジュールが認識した介護情報**： 〈定期的にうがいをしていない〉
ユーザ：お薬の時間は家族が薬を持ってきて知らせてくれます **モジュールが認識した介護情報**： 〈コミュニケーションをとる家族がいる〉 〈服薬のうながし（声かけ）をする家族等がいる〉

これまでにも高齢者を対象にした対話システムの研究開発が行われてきたが [8, 9, 10, 11] 、高齢者による健康状態の報告を認識するために大規模な学習データを構築した例は、我々が知る限り本研究が初めてである。

2 データ構築

まず、アノテータ計 34 名による大規模な学習データの構築を行った。介護質問応答分類モジュールのために 388,691 件の分類つき質問応答ペア、介護情報含意認識モジュールのために 38,868 件の介護情報を含意する発話を作成した。

2.1 介護質問応答分類のためのデータ構築

介護質問応答分類のためのデータ構築では、まず、高齢者の健康状態のチェック項目約 6,000 件の対話シナリオのデータベースから、1,901 件の異なり質問を抽出した。このデータベースは (株) 日本総合研究所が策定を進めているケアマネジメント標準 [1] に基づき作成されたものである。質問はすべて「ストレスはありますか」「塩分制限を守っていますか」のような Yes/No 質問である。

質問は専門用語などを含み、対話システムでそのまま用いるのは適切でない場合があるため、まず、アノテータによる質問の言い換え作業を行った。次に、言い換え後の質問に対して想定される高齢者による応答を作成した。応答の内容に制約はなく、質問に答えていない発話であってもよい。ただし、「はい」「そうです」などの簡単すぎる応答は大規模な学習データを構築しなくても認識可能だと考えられたため、避けるよう指示した。

次にアノテータがそれぞれの応答を YES, NO, UNK, NEGP, OTHER の 5 種類に分類した。YES は、言い方が直接的かどうかにかかわらず「はい」を意味する応答、NO は同様に「いいえ」を意味する応答である。また、UNK はユーザが答えを知らないことがわかる場合である。NEGP は、「かかりつけ医の指示通りに服薬をしていますか」という質問に対する「かかりつけ医なんていないよ」「薬はもらっていません」という応答のように、質問の前提を否定するような応答である。ユーザの応答が NEGP に該当する場合、ユーザに対して適切でない対話シナリオを用いている可能性があり、他と区別して扱う必要があるため、このラベルを設けている。OTHER は上記以外の全てのケースである。具体的には、ユーザが質問に答えず無関係なことを言っている場合や、質問の意図を聞き返している場合などがある。

各応答に対するラベルをアノテータ 3 名が独立に判定し、最終的なラベルは多数決で決定した。3 名とも判断が分かれた質問応答の事例は破棄した。この作業に関わったアノテータは全体で 34 名である。作業を通しての Fleiss' κ [2] は 0.736 であり、高いアノテータ間一致度を示した [5][2]。分類ラベルごとの件数と、応答例を表 2 に示す。

表 2: 質問応答の分類ラベルごとの件数と例

ラベル	件数	応答例
YES	208,088	そりゃちゃんと治したいですから
NO	132,235	つい忘れちゃうんだよね
UNK	6,155	指示通りに飲めているかわかりません
NEGP	3,540	医者なんか行ってないよ
OTHER	38,673	下手なこと言ったら後で怒られるのかな
合計	388,691	

例は「お医者さんの指示通りにお薬を飲んでいますか」という質問に対する応答

2.2 介護情報含意認識のためのデータ構築

介護情報含意認識モジュールは、先述のように、高齢者が自発的に介護情報を提供した場合にそれを認識するためのモジュールである。このモジュールの学習データを構築するため、まず、〈ストレスがある〉〈塩分制限を守っていない〉などの、健康状態を表す文（介護情報と呼ぶ）1,206 件を用意した。介護質問と同様に、介護情報はケアマネジメント標準に基づき作成された対話シナリオから抜き出したものであり、介護質問それぞれについて YES または NO に該当する応答を得たときに得られる情報と同等である[3]。例えば〈ストレスがある〉という介護情報は「ストレスはありますか」という質問に対する YES に対応する。

次に、アノテータ 12 名が、介護情報ごとにそれを含意する発話事例を 3 件作成した。例えば〈ストレスがある〉という介護情報に対して「最近ストレスが多くてね」のような発話を作成した。重複を整理し、38,868 件の発話事例と、38,999 件の発話→介護情報含意関係を得た。

上記の作業の結果得られた発話事例はそれぞれ単一の介護情報と対応することになるが、発話によっては複数の介護情報を含意する可能性もある。例えば「毎日ひとりで買い物に行っていますよ」であれば、同時に〈買い物は自分ひとりでできる〉〈毎日外出している〉〈週に 1 回以上外出している〉などの介護情報を含意している。

このようなケースを効率的に見つけるため、含意関係の推移性を利用することができる。例えば「毎日

[1] https://www.jri.co.jp/page.jsp?id=34346

[2] なお、一部、多数決後に別のアノテータに見直しを依頼したデータや、先にラベルを指定してそのラベルに該当する応答を作成したケースが存在する。ここに示した κ 値は、それらの作業については含めず、当初のアノテーションについて集計した数値である。

[3] 現段階では作業対象を一部の基本的な介護情報に限定しているなどの理由により、介護質問の件数と介護情報の件数は単純に対応しない。

ひとりで買い物に行っていますよ」→〈毎日外出している〉のような発話 U から介護情報 A への含意関係と、〈毎日外出している〉→〈週に 1 回以上外出している〉のような介護情報 A から介護情報 B への含意関係を見つけることができれば、含意関係の推移性 $(U \rightarrow A) \wedge (A \rightarrow B) \Rightarrow (U \rightarrow B)$ を用いることで、「毎日ひとりで買い物に行っていますよ」→〈週に 1 回以上外出している〉の含意関係が得られる。

そこで、まず、類似度の高い文（発話もしくは介護情報）のペアを抽出して含意関係の有無をアノテーションし、その結果に対してさらに含意関係の推移性を用いることで、含意関係の事例を増やすこととした。

類似度の指標として、(i) 内容語のオーバーラップと (ii) トピックワードモデルを用いた指標の 2 種類を使用した。それぞれの指標で類似性が高いと判断されるペアについて含意関係の有無をアノテーションし、最終的にデータをマージした。

(i) の内容語のオーバーラップに関しては、介護情報同士のペアのうち内容語がひとつでも重複するものすべて (33,988 件) を対象に含意関係の有無を判定し、506 件の正例を得た。

(ii) のトピックワードモデルは、ある文 S に対して、その文脈（具体的には、その文および前後の 1 文）に名詞 t が現れる確率 $P(t|S)$ を予測する BERT ベースのモデルである。これを用いて、文の類似度を以下のように見積もる。文のペア $\langle S_1, S_2 \rangle$ について、その類似度を S_1 が S_2 と同じ文脈に生起する確率 $P(S_1|S_2)$ として定義する。$P(S_1|S_2)$ は、トピックワードモデルを用いて以下のように見積もることができる。

$$
\begin{aligned}
P(S_1|S_2) &\approx \sum_t P(S_1|t)P(t|S_2) \\
&= \sum_t \frac{P(t|S_1)P(S_1)P(t|S_2)}{P(t)}
\end{aligned}
$$

ここで $P(S)$ の一様分布を仮定すると以下を得る。

$$
P(S_1|S_2) \propto \sum_t \frac{P(t|S_1)P(t|S_2)}{P(t)}
$$

この右辺を類似度の尺度として利用した。

トピックワードモデルで予測対象となる名詞は、我々のチームで開発している質問応答システムである WISDOM X [6] のデータベースにおける高頻度名詞上位 1 万語とした。学習時の入力はコーパスから抽出した文、出力はその文もしくはその前後の文に現れる名詞（または、予測対象となる名詞が一つもないことを表す特別なシンボル）のリストである。

トピックワードモデルに基づく類似度の指標に基づき、発話同士、発話→介護情報、介護情報同士に対して、それぞれ類似度の高いペア上位 50 万件から 10 万件、10 万件、5 万件をランダムサンプリングしたものを対象として含意関係の有無をアノテーションし、それぞれ 150 件、5,279 件、5,030 件の正例を得た。

(i) 内容語のオーバーラップを利用したアノテーション手法で得られた正例と (ii) トピックワードモデルを利用したアノテーション手法で得られた正例、および先に発話作成作業によって得られていた正例をマージし、さらに推移性を用いて正例を拡張した。推移性によるデータの拡張は、それ以上新しい含意関係が見つからなくなるまで再帰的に適用した。最終的に、発話→介護情報含意関係の正例 76,910 件が得られた。

3 モジュールの学習と評価

本節では、前節で構築したデータを用いたニューラルネットワークによるモジュールの実装について述べる。介護質問応答分類モジュールと介護情報含意認識モジュールは、どちらも独自に事前学習を行った BERT [1] を fine-tuning する形で実装した。BERT の事前学習には、Kadowaki ら [4] と同様に CRF ベースの因果関係認識 [7] で因果関係ありと判定されたパッセージをコーパスとして利用した。介護質問応答分類モジュールは $\text{BERT}_{\text{LARGE}}$ を用い、22 億文をバッチサイズ 4,096 で事前学習した。介護情報含意認識モジュールは $\text{BERT}_{\text{BASE}}$ を用い、4 億文をバッチサイズ 1,024 で事前学習した。

3.1 介護質問応答分類の評価実験

介護質問応答分類モジュールは、質問と応答のペアを特別なトークンである `[SEP]` を区切りとして連結し、別々のセグメント埋め込みを割り当てたものを入力とし、YES, NO, UNK, NEGP, OTHER のどれかのラベルを出力とするモデルである。

同じ元質問から作成された質問応答は同一のセットに入るようデータを分割して実験を行った。データサイズは表 3 に示した。

表 3: 介護質問応答分類モジュールの実験データサイズ

Train	Val	Dev	Test	計
285,696	35,520	32,982	34,493	388,691

ハイパーパラメータは学習率 1e-5, 2e-5, 3e-5, 4e-5, 5e-5、エポック数 1, 2, 3, 5, 10, 20 の組み合わせを探索した。ここでは開発セット (Dev) で最高性能になったモデルの評価セット (Test) での性能を報告する。性能は分類ごとの平均精度 (AP) のマクロ平均で評価した。結果を表 4 にまとめた。データ作成において「はい」や「そうです」などの簡単な応答を抑制したにもかかわらず、YES では 97.3%、NO では 94.1%という平均精度を実現している。事例数が少ない NEGP が低い

精度にとどまっており、今後の改善が必要である。全分類を通じての平均精度のマクロ平均は 89.0%であった。正解率で見ると全体で 88.4%であった。

表 4: 介護質問応答分類モジュールの平均精度

YES	NO	UNK	NEGP	OTHER	マクロ平均
97.3	94.1	88.2	77.2	88.5	89.0

3.2 介護情報含意認識の評価実験

介護情報含意認識モジュールは、発話が 1 件入力されるごとに、1,206 の介護情報についてそれぞれ含意の有無を 2 値分類する必要がある。これを 1,206 個の別々の出力をもつマルチラベルで行う手法（multi-label）と、それぞれの介護情報を発話と連結して入力とし、出力は 1 個とする手法（two-segment）の 2 種類を試した。後者の手法では、介護質問応答分類モジュールと同様、介護情報と発話を特別なトークンである `[SEP]` を区切りとして連結して入力とした。この手法は、予測時に介護情報を自由に追加でき、介護情報の内容も学習・予測に使用できるなどのメリットがある一方、学習・予測ともに時間がかかるというデメリットがある。

実験では、人手で正例と判断されたもの、また推移性を利用して正例と判断できたものを除き、それ以外のすべての発話→介護情報ペアをデフォルトで負例扱いとした。このため、正例 76,910 件に対し、負例が 46,797,898 件となり、負例が圧倒的に多いデータとなっている。表 5 に実験に用いるセットごとのデータの規模を発話数で示した（データポイントの数は、これに介護情報の数である 1,206 を掛けた数となる）。

表 5: 介護情報含意認識モジュールの実験データサイズ

Train	Val	Dev	Test	計
29,891	2,989	2,990	2,998	38,868

ハイパーパラメータは学習率 1e-5, 2e-5, 3e-5, 4e-5, 5e-5、エポック数は multi-label では 20, 50, 100, 200、two-column では 1, 2, 3, 5 を探索した。モデルの性能は介護情報ごとの平均精度のマクロ平均で評価し、開発セット (Dev) で最も高い性能を示したモデルの評価セット (Test) での性能を示す。実験結果を表 6 にまとめる。結果は two-segment で 87.1%となり、multi-label の 82.9%を上回った。

表 6: 介護情報含意認識モジュールの平均精度

デザイン	マクロ平均
Multi-label	82.9
Two-segment	**87.1**

4 おわりに

本研究では、高齢者の健康情報をモニタリングする対話システムの実現に向けて開発した 2 つのモジュールについて報告した。

データ作成はこれまでアノテータに依頼する形で進めてきたが、アノテータと実際に介護用対話システムのユーザとなる高齢者との間にはさまざまな面で違いがあることが予想される [3]。現在、高齢者を対象とした実証実験がすでに実施されており、今後はその結果も踏まえてモジュールの改善を行っていく予定である。

謝辞 本研究は総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム (SIP)[4]「Web 等に存在するビッグデータと応用分野特化型対話シナリオを用いたハイブリッド型マルチモーダル音声対話システムの研究」（管理法人：JST）によって実施されたものである。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL: HLT 2019*, pp. 4171–4186, 2019.

[2] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382, 1971.

[3] Kallirroi Georgila, Maria Wolters, Johanna D Moore, and Robert H Logie. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, Vol. 44, No. 3, pp. 221–261, 2010.

[4] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proc. of EMNLP-IJCNLP 2019*, pp. 5815–5821, 2019.

[5] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.

[6] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In *Proc. of COLING 2016 (Demo)*, 2016.

[7] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *Proc. of ACL 2013*, pp. 1733–1743, 2013.

[8] 大竹裕也, 萩原将文. 高齢者のための発話意図を考慮した対話システム. 日本感性工学会論文誌, Vol. 11, No. 2, pp. 207–214, 2012.

[9] 鶴田直之, 重田義和, 前田佐嘉志, 高橋伸也, 森元逞. 在宅健康管理システムのための対話システム. ヒューマンインタフェース学会研究報告集, Vol. 4, No. 4, pp. 25–29, 2002.

[10] Maria Klara Wolters, Fiona Kelly, and Jonathan Kilgour. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. *Health Informatics Journal*, Vol. 22, No. 4, pp. 854–866, 2016.

[11] 山本大介, 小林優佳, 横山祥恵, 土井美和子. 高齢者対話インタフェース：『話し相手』となって、お年寄りの生活を豊かに. 電子情報通信学会技術研究報告. HCS, ヒューマンコミュニケーション基礎, Vol. 109, No. 224, pp. 47–51, 2009.

[4] https://www8.cao.go.jp/cstp/gaiyo/sip/