E-024

高齢者対話インタフェース
―発話間の共起性を利用した傾聴対話の基礎検討―

Development of Dialog Interface for Elderly People
-Active Listening with Word Co-occurrence Analysis between Utterance Pairs-

小林　優佳†　　山本　大介†　　土井美和子†
Yuka Kobayashi　Daisuke Yamamo　Miwako Doi

1.はじめに

日本では人口の 22.7%が 65 歳以上であり、そのうちの 22%は独居高齢者であり[１]、孤独死や認知症治療が問題となっている。人との対話コミュニケーションがこれらの予防に有効と言われており、独居高齢者の住宅や介護施設に出向いて高齢者の話し相手を行う”傾聴ボランティア”が注目されている[２]。傾聴とは高齢者の話を「聴く」ことで精神安定を図るものである。しかし、介護分野の人材不足により十分な話相手が得られていない。

このような現状に対し、我々は卓上インタフェースロボット ApriPoco^{TM}(図 1)を用い、高齢者の話の”聴き手”となる対話インタフェースの研究開発に取り組んでいる[２]。

話の「聴き手」は相手の話を理解し、それを伝えるための発話が必要である。本システムでは音声認識を行い、認識結果に基づいた発話を行う。しかし音声認識には認識誤りが含まれる。カーナビのようなタスク志向型対話とは異なり、ユーザの発話全てを聞きとる必要はないが、ユーザが発話していない内容を誤認識して発話するのは好ましくない。

音声認識では音響モデルと言語モデルによって正解が求められる。言語モデルは N-gram を用いるものが多い。N-gram は単語の生起確率を直前の N-1 個の連続する単語から推定するものである。計算機の処理能力などの制限から N=3 を用いることが多い。順序制約が大きく、カーナビや音声検索への入力のように文法的に正しく発話されやすい発話に対する性能は高いが、雑談のように言い淀み、言い直し、倒置が発生する発話には弱い。また、N 個の単語しか考慮しないため局所的な文脈しか考慮せず、長距離文脈が考慮されない。

そこで、N-gram の欠点を補い、認識精度をあげるための手法として LSA(Latent Semantic Analysis)が提案されている。これは認識結果中の単語間の共起性に注目したものである。あらかじめ複数の文書を含むコーパスを使用して任意の二つの単語について同じ文書で使用される共起率を求めておく。そして、認識結果中の任意の二つの単語に対して共起率を求め、生起確率を求める[４][５][６]。

しかし、音声認識結果中の単語間の共起性で判定すると、誤認識単語同士で共起する場合があり、誤って判定される危険がある。そこで、音声認識結果中ではなく、システムの発話・音声認識結果間の共起性によって誤認識単語を排除する手法を提案する。

雑談では対話参加者は必ず対象となる話題に関する発話を行うため、発話間には意味的な共起性がある。システムの直前の発話中の単語と共起性が低い音声認識結果中の単語は誤認識の可能性が高いので排除することで、誤認識した単語による検討はずれな発話を防ぐことができる。

本稿ではこの「発話間共起フィルタ」を使用した傾聴モードの対話戦略について述べ、その性能について報告する。

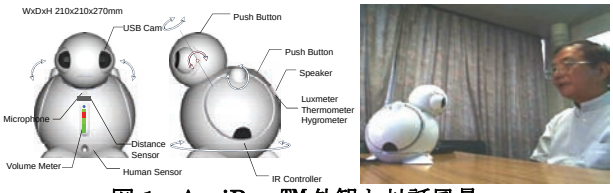

図 1　ApriPoco^{TM}外観と対話風景

2.発話間共起フィルタ

2.1 発話間共起フィルタ概要

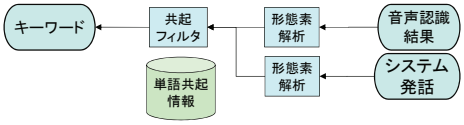

図 2　発話間共起フィルタシステム構成図

図 2は発話間共起フィルタのシステム構成図である。本システムではシステムとユーザは交互に発話を行う。ユーザ発話が終了するとシステムはユーザの発話を音声認識する。

システムは自分自身が直前にした発話と、音声認識結果をそれぞれ形態素解析し、名詞、形容詞、動詞、形容動詞（自立語）を抽出する。助詞、助動詞は話題に関する意味を持たないため、意味的な共起性の判断ができないので除外する。音声認識結果 r_m 中の自立語 $w_i(r_m)$に対して下記を算出する。

$$S_p(w_i(r_m)) = \max_j \left\{ C(w_i(r_m), w_j(u_{m-1})) \right\} \qquad \text{Eq 1}$$

$w_j(u_{m-1})$はユーザ発話の一つ前の発話、つまりシステム発話中の単語を表す。$C(w_i(r_m), w_j(u_{m-1}))$は $w_i(r_m)$と $w_j(u_{m-1})$の単語ペアの共起率を表す。システム発話中の単語 $w_j(u_{m-1})$に対してそれぞれ共起率を求め、最大値を $w_i(r_m)$の共起スコア $S_p(w_i(r_m))$とする。

システムは各自立語 $w_i(r_m)$に対して $S_p(w_i(r_m))$を算出し、$S_p(w_i(r_m))$の高い順に優先的に次の発話に使用する。

次に共起率の算出方法について説明する。単語間の共起性の算出方法は相互情報量、情報利得、対数尤度比[７]などが考えられるが、ここでは事前検討で最適な結果を出したコサイン係数を使用する。コサイン係数は以下で算出される。

$$C_c(w_A, w_B) = d_{AB} / \sqrt{f_A f_B} \qquad \text{Eq 2}$$

†（株）東芝　研究開発センター, Corporate Research & Development Center, Toshiba Corporation

f_A: 単語 A の頻度
f_B: 単語 B の頻度
d_{AB}: 単語 A と単語 B を含む文書数

共起率算出の元になるテキスト情報には Google™ n-gram[８]を使用した。これは Google™ が Web から抽出した約 200 億文の文章から生成されたデータである。7-gram を使用し、7-gram を 1 つの文書とみなして算出した。各 7-gram から自立語を抽出し、単語の頻度 f_A/f_B と d_{AB} を算出した。5.8 億個の 7-gram から 215 万個の自立語と 9311 万個の単語共起情報が生成された。この単語共起情報を元に $C(w_i(r_m), w_j(u_{m-1}))$を算出する。

2.2 同音異義語の解決

音声認識では音声で入力された文章に対して漢字を割り当てる必要がある。単語間の共起性を使用することで同音異議語に対して正しい漢字を割り当てることができる[９][１０]。そこで、$S_p(w_i(r_m))$の算出方法を以下のように変更する。

$$S_p(w_i(r_m)) = \max_j \{C(pron(w_i(r_m)), w_j(u_{m-1}))\} \quad \text{Eq 3}$$

$pron(w)$は単語 w の読み仮名を表す。システム発話中の単語 $w_j(u_{m-1})$は正しい漢字が割り当てられているので、こちらは漢字で検索する。あらかじめ単語共起情報には単語の読み仮名、品詞などを付与しておく。単語 A の読み仮名が $pron(w_i(r_m))$と一致し、単語 B が $w_j(u_{m-1})$と一致するような単語ペアを検索することで、$w_i(r_m)$に最も適切な漢字を割り当てることができる。

2.3 一般語による性能悪化

自立語の中にも内容的な意味を持たず、文法的な使われ方をする一般語と呼ばれるものがある。「する」「なる」「ある」「いる」などがこれにあたる。これらは文章の内容に関係なく頻出するため、高い共起率を持つ。しかし、意味を持たない単語との共起性で判定してしまうと誤りの可能性が高い。そのため、一般語を含む共起情報は削除する必要がある。これに対して内容的に意味を持つ単語を内容語と呼ぶ。

一般語は情報検索では検索結果に悪影響を及ぼすので、検索キーから除去するために除去方法が提案されている[１１][１２]。我々はその中で IDF(Inverse Document Frequency)を使用した一般語除去を用いる。単語 w の IDF は以下で算出される。

$$IDF(w) = \log(|D|/d) \quad \text{Eq 4}$$

|D| : 全文書(7-gram)数
d : 単語 w を含む文書(7-gram)数

IDF は単語の一般性を表す指標である。IDF が低いほど一般性が高い。IDF の低い単語を一般語とみなし、IDF を含む共起は共起性の判断に使用しない。$w_i(r_m)$の IDF は以下で算出される。

$$I_p(w_i(r_m)) = \min\{IDF(w_i), IDF(w_b)\}$$

$$w_b = \arg\max_{wj} \{C(pron(w_i(r_m)), w_j(u_{m-1}))\} \quad \text{Eq 5}$$

共起スコア S_pを求める際に最大共起率をとった単語ペアの $w_j(u_{m-1})$の IDF と $w_i(r_m)$の IDF のうち、値が小さい方を $w_i(r_m)$の IDF とする。$w_i(r_m)$と最も共起する単語 $w_j(u_{m-1})$が一般語だった場合、$w_i(r_m)$は内容語との共起性が認められないと判断され、正解から除外される。

このように IDF が低い単語による共起性を除去することで一般語との共起による誤判定を防ぐことができる。

3. 発話間共起フィルタを用いた音声対話システム

発話間共起フィルタを使用したシステムについて説明する。図 3はシステム構成図である。

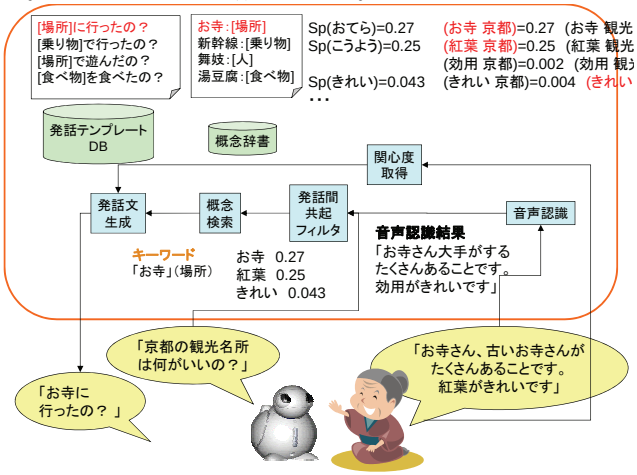

図 3　音声対話システム　システム構成図

ユーザの話の「聴き手」となるロボットは、ユーザに聞いていることを伝えるために発話を行う。ここではカウンセラーの発話方法を参考にし、単純相槌、反復相槌、質問の 3 種類の発話を行う[１３][１４]。単純相槌は「へえ」「うんうん」などの言語情報を含まない相槌である。反復相槌は「へえ、京都ね」「友達とね」のように相手の発話内容の一部をそのまま繰り返す相槌である。質問は話を広げるために相手の話に対する質問を行う。

どの発話をするべきかは関心度によって判断する。関心度とはユーザが話題に興味をもっているかどうかの指標であり、表情・韻律から判定される[１５]。ロボット・ユーザが対話をしている動画を第三者が見て評価した値を関心度の正解値とし、表情・韻律から自動判定したところ、4 段階で 76%の正答率だった。システムはユーザの関心度が高ければ単純相槌・反復相槌を行い、低ければ質問を行う。

質問文生成には概念辞書と発話文テンプレートＤＢを使用する。概念辞書は単語と単語の概念をあわせて格納した辞書である。発話文テンプレートは発話文の一部が概念に置き換えられたもので、ここに該当する概念の単語を入れることで発話文を作成することができる。

このシステムを使用した発話方法について説明する。システムはまず、定型の文章を発話する。最初以外にも、発話文生成に失敗した場合は定型文章を発話する。そしてユーザの発話内容を音声認識する。認識結果・システム発話文章を発話間共起フィルタに入力し、共起スコア S_p とともに自立語が出力される。図 3の例では認識結果は「効用」だったが、「京都」と共起性が高い「紅葉」に変更され、誤認識が修正されている。一番共起率が高い自立語をキーワードとして取得する。ここでは、キーワード「お寺」が選択される。反復相槌を生成する場合は「お寺なんだね」という発話を行い、質問文を生成する場合は、キーワードを概念辞書で検索し、概念[場所]を取得する。発話テンプレート DB で概念[場所]を使用する発話文テンプレートを検索し、見つかればキーワードと発話文テンプレート

（[場所]に行ったの？）を組み合わせて「お寺に行ったの？」という文章を次に発話する。

4. 実験 1

4.1. 実験用データ

発話間共起フィルタの性能をオフラインで評価するために、ロボットは固定の質問文を発話し、被験者がそれに応えるという対話実験を行った。被験者は初見の 22 人（60 代男女 6 人ずつ、20 代男女 5 人ずつ）で、京都旅行について 4 回の対話を行った。被験者は接話マイクをつけて音声収録を行った。

録音に失敗したものなどを除いて 83 対話のデータが収集され、平均ユーザ発話回数は 18 回だった。

4.2 共起フィルタの閾値の決定

共起率スコアがあまりに低い共起は検出されても誤認識である可能性が高い。そのため、共起率スコアに閾値を設ける。この閾値を決定する。

本システムではユーザの発話を全て聞き取る必要はないが、ユーザが発話していない単語を誤認識してその単語を使用して発話を行うのは好ましくない。そのため、再現率（＝正認識単語数/発話単語数）よりも適合率（＝正認識単語数/認識単語数）が重要である。

実験データの音声認識結果に対して発話間共起フィルタを適用し、閾値を決定する。閾値に対して適合率・再現率の変化は図 4のようになった。

最も適合率の高い箇所を閾値として選択してすると再現率が 16%、適合率 61%となった。

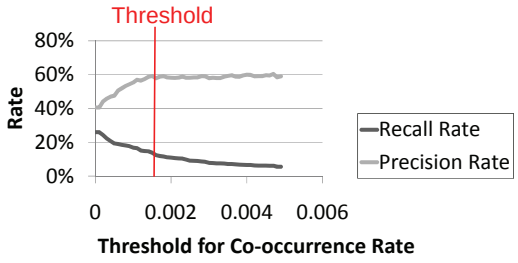

図 4　共起フィルタの閾値と適合率の変化

4.3 発話内共起性との比較

次に、発話間共起性と音声認識結果内の共起性（発話内共起性）の性能比較を行う。音声認識結果内の自立語の共起性は以下の式で算出される。

$$S_s(w_i(r_m)) = \max_j \{C(pron(w_i(r_m)), pron(w_j(r_m)))\} \quad \text{Eq 6}$$

発話間共起 S_p と音声認識結果内共起 S_s を比較したものが図 5である。

発話間共起フィルタの方が、適合率が高いことがわかる。誤認識結果同士が共起することが問題であることが考えられる。ただし、音声認識結果内で単語ペアを生成すると 1 つの単語ペアから 2 個の単語が取得できるので、出力できる単語数は多く、再現率は高い。

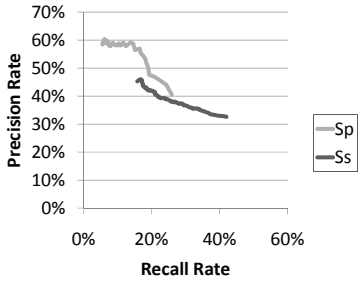

図 5　発話間と音声認識結果内共起性の比較

4.4 IDF フィルタによる性能評価

次に、IDF が低い単語による共起性を除去することで一般語による判定誤りを防ぐことが可能か検証を行う。4.2 で求めた共起スコアの閾値で判定を行い、正しい認識と判断された単語に対して I_p を求め、I_p に対して閾値を設定し、閾値未満の単語は誤認識として除外する。I_p の閾値を変化させると適合率・再現率の変化は図 6のようになった。

最も適合率の高い箇所を閾値として選択してすると再現率が 6%、適合率 66%となり、4.2 の結果から改善されていることがわかる。

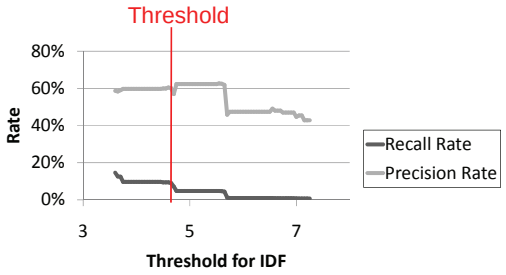

図 6　IDF フィルタの閾値と適合率の変化

4.3 クロステスト

実験データの 83 個の対話を 9 グループに分割し、8 グループで閾値を決定し、残りの 1 グループに適用し、性能評価を行うクロステストを行った。

表 1　クロステスト結果

	発話間共起フィルタ			音声認識	
グループ	適合率	再現率	単語取得率	適合率	再現率
1-9	100%	11%	9%	68%	34%
10-18	73%	9%	13%	39%	29%
19-27	46%	4%	5%	34%	20%
28-36	83%	3%	3%	27%	9%
37-45	51%	4%	11%	38%	27%
46-54	79%	10%	16%	64%	43%
55-63	73%	5%	11%	37%	27%
64-72	65%	3%	8%	38%	24%
73-83	72%	7%	12%	45%	33%
average	71%	6%	10%	43%	27%

表 1は評価結果である。単語取得率は正解単語数/ユーザ発話文数で算出されている。元の音声認識結果では適合率 43%だったのが 71%まで改善されたのがわかる。ただし、単語取得率 10%ということはユーザ発話 10 回に対して 1 個しか単語が取得できない。つまり、残りの 9 回は認識結果を使用できないので、固定文を発話するか、過去に取得した単語で複数回発話する必要がある。

5. 実験２

ここまでは音声認識の性能評価を行ったが、ここでは音声対話システムとしての性能評価を行う。発話間共起フィルタによって出力された単語を使用して反復相槌を生成し、3 人の評価者によってユーザの発話に対して適切かどうかの判定を行った。

具体的には以下のようなシステム発話、ユーザ発話、生成した反復相槌の連続する 3 つの発話を評価者に見せ、ユーザの発話に対して反復相槌が「非常に適切である(5)」「適切である(4)」「わからない(3)」「適切ではない(2)」「全く適切ではない(1)」の 5 段階のどれに当たるかを評価してもらった。3 人の評価結果を点数ごとに分類したものが表 2である。

システム：京都では何がお勧め？
ユーザ：観光するのがいいですよ。
システム：観光なんだね

表 2　反復相槌性能評価

点数	割合 (数)
1-2	31% (218)
3	16% (113)
4-5	53% (377)

生成された相槌の 53%は適切であると判定されていることがわかる。点数が 1,2 であった相槌の内容を見たところ、不適切と判定されたものには大きく 2 種類の要因があることがわかった。

一つ目は、発話間共起フィルタは音声認識の挿入誤りを改善するものであるが、削除誤りには対応できないことである。認識されなかった単語は相槌に使われることはないので、相槌に適切な単語が認識されなかった場合には不適切と判定される。

二つ目は、時制と肯定・否定の不一致である。自立語を抽出する際に動詞・形容詞・形容動詞は基本形に直して抽出するので、相槌に使われる際も基本形のまま使用される。すると下記のようにユーザの発話した事実と異なる相槌が発話される。

ユーザ：今日は朝から何も食べてないよ
システム：食べたんだね

システムは活用のあるものに対しては時制、肯定・否定を一致させて発話する必要がある。しかし、音声認識で使用されていた時制をそのまま使用すると誤認識の可能性がある。

6. まとめ

高齢者の話の聴き手になる音声対話ロボットを実現するため、発話間共起フィルタを提案した。発話間共起フィルタはシステム発話中の単語とユーザ発話の音声認識結果中の単語の共起性を計測し、共起性の高い単語を使用することで誤認識を排除する手法である。この手法を実験データに適用したところ、音声認識では適合率 43%であったのが適合率 71%と 28%の改善が見られた。再現率は大幅に減少するが、対話では全ての単語を聞き取れる必要はなく、聞き取れた単語に基づいて発話を行えばよいので、再現率の減少はあまり問題にはならない。

この手法を用いて反復相槌を生成し、性能評価をしたところ、53%の相槌は適切であるという評価結果になった。不適切と判定された相槌は音声認識の削除誤り、時制、肯定・否定の不一致が原因であることがわかった。

本システムで用いた共起情報は 1 つの文章から算出されたものであるが、複数の発話間の共起性を判断にすることにも適していることがわかった。

今後はこの手法を用いて実際に対話実験を行い、よりよい音声対話システムの実現を目指す。

謝辞

本研究の一部は総務省の委託研究により実施したものである。

参考文献

[1] 内閣府, 高齢社会白書平成 22 年版, (2010)
[2] NPO 法人ホールファミリーケア協会: 新傾聴ボランティアのすすめ, 三省堂, (2009)
[3] 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ー『話し相手』となって、お年寄りの生活を豊かに。ー, HCS2009-56, pp. 47-51, (2009)
[4] S. Cox, S. Dasmahapatra: High-level approaches to confidence estimation in speech recognition, *Speech and Audio Processing, IEEE Transactions on* , vol.10, no.7, pp. 460- 471, (2002).
[5] Q. Gao, X. Lin and X. Wu: Just-in-time latent semantic adaptation of language model for Chinese speech recognition using web data, *Spoken Language Technology Workshop, IEEE.*, (2006).
[6] J. R. Bellegarda: Large Vocabulary Speech Recognition with Multispan Statistical Language Models, *Speech and Audio Processing, IEEE Trans-actions,* vol.8, no.1, pp.76-84, (2000).
[7] T.Dunning: Accurate methods for the statistics of surprise and coincidence, *Computational linguistics*, vol.19, no.1, pp.61-74 (1993).
[8] GSK2007-C, Web 日本語 N-gram 第 1 版, 2007, http://www.gsk.or.jp/catalog.html
[9] 元永靖和, 池原悟, 村上仁一: 動詞と名詞の意味的共起関係を用いた同音異義語のかな漢字変換, NLC, 電子情報通信学会, vol.99, no.228, pp.17-24 (1999)
[10] 鳥原信一: テキストの共起単語情報を用いたかな漢字変換, 情報処理学会全国大会, pp.225-226, (1993)
[11] 間瀬久雄, 大西昇:特許文書中のタームの出願人別使用傾向の分析と類似特許文書検索精度への影響評価, DD, 情報処理学会, no.33, pp.77-84 (2002)
[12] 岡田真, 浜田浩史, 宝珍輝尚: マルチメディアデータの効率的検索のためのキーワード自動抽出手法, 自然言語処理研究会研究報告, 情報処理学会, no.94, pp.73-78 (2005)
[13] 宮崎聡子, 上図な聴き方が身につく技術, あさ出版, (2008).
[14] A. E. Ivy, M. B. Ivey and C. P. Zalaquett: International Interviewing & Counseling Facilitating Client Development in a Multicultural Society, CA, USA: Brooks/Cole Pub Co. (2009).
[15] 小林優佳, 山本大介, 横山祥恵, 土井美和子: 高齢者向け対話インタフェース―雑談時における関心度検出方法と関心度を利用した音声対話インタフェース―, SIG-SLUD, 人工知能学会, no.59, pp.1-6 (2010)