連想システムを用いた話者理解

八尾 旬扇,　渡部　広一,　河岡　司
同志社大学　工学部　知識工学科

１．はじめに

会話には，話者の好みが関係していることが多く，それをコンピュータが理解し会話に使用することで，より人間同士のコミュニケーションに近づくことができると考えられる．そこで，人間とコンピュータがコミュニケーションしていく上でも，誰と話しているかを認識し，発言の中から話している相手の特徴や好みなどの個人情報を取得する話者理解が必要になる．

本研究では，入力された文の中から個性や個人の好みに関するキーワードを抽出し，連想を用いてそれを拡張する(関連語を増やす)ことで個人の好みや特徴に関する重要な情報を効率よく取得する方法を提案する．また座標を使うことで，血縁関係・友人関係などの人間関係を理解させ，それを考慮して情報を体系的に取得している．

２．話者理解

本研究では，誰と話しているかを認識し，発言から話している相手の特徴や好みなどの個人情報を取得することを話者理解としている．話者理解を行う際，まず誰と話しているかを認識するために，あらかじめ名前を入力することで話者を特定する．また，相手の特徴や好みなどの個人情報を格納する場所として，あらかじめ多くの人が持っていると思われる項目を 47 項目設定した．この項目は，誰もが持っている一般的な情報（名前・性別・生年月日・血液型・星座・身長・体重・家族構成・職業・出身地・現住所・国籍・利き手などの 21 個），個性に関わる情報（性格・長所・短所・趣味・特技・持病・チャームポイントの７個），好みに関する情報（好きな色・嫌いな色・好きな食べ物・好きなスポーツ・嫌いなスポーツ・好きなキャラクター・ファン・行ってみたい国などの 17 個），最後に会った時間，人間関係から成り立っている．本研究の主な目的は，入力文からこれらの項目に当てはまる情報を取得することと，家族関係・友人関係などの人間関係を理解させることである．一人の人物の情報・項目の取得例を表１に示す．

表１：話者理解の例

項目	情報	項目	情報
名前	山本優子	趣味	音楽鑑賞
ふりがな	やまもとゆうこ	好きな色	ピンク
性別	女	好きなスポーツ	テニス
利き手	右	嫌いな食べ物	
血液型	A	好きな食べ物	
職業	大学生	性格	明るい
ペット		短所	頑固

次に，処理の流れを説明する．まず，前回までに取得した情報を読み込む．名前がある場合は，前回からの期間に応じて挨拶を行う．一方，名前がない場合は，初対面で聞いてもおかしくない項目や外見でわかる項目の情報（出身地，名前の漢字表記，性別，利き手，現住所）を取得する．そして，情報フレーム[1]に分けられた形で文を入力し，パターン選別を行う．情報フレームとは，「主体」「何」「時間」「場所」「理由」「方法」「誰に」「用言」の６W１Hからなるフレームのことである．パターン選別については，４章で説明する．

次に，入力文から情報と項目を取得するが，ここでは形態素解析ソフト「茶筌[2]」で解析し，名詞や形容詞を抽出した後，シソーラス[3]を使って拡張し，項目との関連度計算[4]で精錬を行う．シソーラス・関連度計算については，３章で説明する．以上が，話者理解メカニズムの大まかな流れになっている．

３．概念ベース・関連度計算・シソーラス

３．１　概念ベース

概念ベース[5]とは，ある語(概念)とそれを表す単語集合(属性)から構成されている．この概念と属性のセットは約９万語登録されている．また，この概念と属性のセットにはその重要性をあらわす重みが付与される．

３．２　関連度計算

関連度[4]とは，二つの概念AとBの関連の強さを定量的に評価するものである．関連度は，０から１までの連続値をとり，概念同士の関連が大きい時は高い値となり，関連が小さい時は低い値となる．

３．３　シソーラス

シソーラス[3]とは一般名詞を整理したもので，約 2700 の意味属性の上位下位関係・全体部分関係が木構造で示されたものである．約 13 万語が登録されており，意味属性であるノードとノードに属している単語であるリーフから構成されている．また，親子・兄弟関係を持つ語についてはその関係も記述している．

４．パターン選別

話者理解とは，相手の個性や好みに関する情報を取得することであるが，会話の中に出てくるすべての人の情報を取得しても，自分の知らない人や，直接関係のない人（例えば，祖母の友達，先輩の母など）の情報は，その後の会話に使われにくい．そこで，主体の主語と修飾語によって，自分に関係があると考えられる５つのパターンを作成した．

(1)主語が一人称，または主語がない時
(2)修飾語が一人称で,主語が「物」の時　[例]私の犬が…
(3)主語が固有名詞の時　[例]山田さんは…
(4)主語が「彼」,「彼女」などの時
(5)修飾語が一人称で主語が「人」の時　[例]私の母は…

本研究では，この(１)～(５)のいずれかに当てはまる時のみ，入力文中の個人情報を取得することにする．情報の取得方法は各パターンで異なるため，それぞれについて５章で説明する

５．個人データ・項目取得

一般的に，個人に関する重要な情報は「何」フレームに含まれていることが多い．そこで，「何」フレームを中心に項目に当てはまる情報を取得することを考える．

５．１　個人データ・項目取得方法 ～パターン(1)～

項目候補の取得と選定処理の２つが主な処理である．

(1)項目候補の取得
①「何」フレームに入力があるかどうかを調べる．
②情報の取得：入力がある場合は，形態素解析ソフト「茶筌」を用いてその中の名詞Ａを取り出す．これが，取得する情報となる．
③親ノードの取得：項目名は具体的な名詞より，何かの集合を表す抽象的な名詞が多い．そこで，先程「何」フレームから取得した名詞Ａの直接の親ノードＢを取得する．
④項目候補の取得：A, B とすべての項目の関連度を計算し，それぞれの名詞について高関連度の項目を４つずつ取得する．これが，名詞Ａが入る項目の候補となる．

（2）選定処理
①動詞の分類：動詞を「好き」に関する語，「嫌い」に関する

語，普通の動詞の３種類に場合分けする．

②項目の削除：項目に条件をつけ，候補の中でその条件に当てはまらないものを削除していく．

③②の処理が行われた後に残っている項目を最終的な項目候補とする．

一方，「何」フレームに入力がない場合は，「用言」フレームに注目する．「用言」フレームに形容詞・形容動詞が入っている場合は，話者の性格を表していることが多い．そこで，「用言」フレームに入っている形容詞や形容動詞を「性格」とする．具体例を図１に示す．

[例１]私はテニスが好きだ．
①「何」フレーム→「テニス」
②「テニス」のシソーラス直親ノード→「スポーツ」
③項目との関連度計算
④「テニス」→チャーム・ポイント,好きなスポーツ,嫌いなスポーツ,ファン
「スポーツ」→好きなスポーツ，嫌いなスポーツ，ファン，体重
⇒項目候補：チャーム・ポイント,好きなスポーツ,嫌いなスポーツ,ファン,体重
⑤選定処理を行う
⑥情報：「テニス」，項目：「好きなスポーツ」となる．

図１：パターン（１）の例

５．２　個人データ・項目取得方法 ～パターン(２)～

修飾語が一人称で主語が「物」の場合は，「何」フレームの前に主語に注目する．それは，主語が項目と一致していれば，項目がすぐに決定されてしまうので情報のみ取得することを考えればよいが，主語と項目に一致するものが無ければ，パターン(１)と同様，情報・項目とも取得することを考える必要があるからである．

そのため，まず初めに主語と一致する項目があるかどうかを調べる．主語と一致する項目が見つかった場合は，「何」フレームに入力があるかどうかを調べる．入力がある場合，「何」フレームの中身が取得情報となり，「何」フレームに入力がない場合は，「用言」フレームの形容詞・形容動詞を「性格」とする．

一方，主語と一致する項目がない場合は，主語と項目の関連度計算を行い，高関連度の項目を４つ取得し，項目の候補とする．次に，選定処理を行う．方法は，パターン(１)の時と同じである．具体例を図２に示す．

[例１]私の趣味は音楽鑑賞です．
①主語→「趣味」　⇒　項目「趣味」と一致する．
②「何」フレーム→「音楽鑑賞」
③情報：「音楽鑑賞」，項目：「趣味」となる．
[例２]私の誕生日は４月 10 日だ．
①主語→「誕生日」　⇒　一致項目なし
②「誕生日」と項目の関連度計算を行う．
③選定処理を行う　⇒　項目→「生年月日」となる
④「何」フレーム→「４月 10 日」
⑤情報：「４月 10 日」，項目：「生年月日」となる．

図２：パターン（２）の例

５．３　個人データ・項目取得方法 ～パターン(３)～

まず初めに，すでに持っているデータファイルの中と主語の固有名詞が一致する人が何人いるのかを調べる．一人のみの場合は，その人が情報取得対象者となり，次に情報・項目取得の操作に移る．この場合の情報と項目の取得については，パターン（１）の操作と全く同じである．

一方，データファイルに主語の固有名詞と一致する人が二人以上いる場合は，話者の人間関係を利用し，誰のことを話しているのかを特定する．人間関係の表し方については６章で述べる．この時，以下のような状況が考えられる．

①情報取得対象者が一人に絞れた場合
パターン(1)と同じ操作を行い，情報と項目を取得する．

②情報取得対象者が複数のままである場合
対象者の名前を提示し，「どの人のことですか？」と聞き返す．また，同時にパターン(1)と同じ操作を行い，情報と項目も取得しておく．

③情報取得対象者が見つからなかった場合
「○○さんとは誰ですか？」と聞き返す．この場合も，情報と項目は他と同様の操作を行い取得しておく．

以上が，主語が固有名詞の場合の処理の流れになっている．具体例を図３に示す．

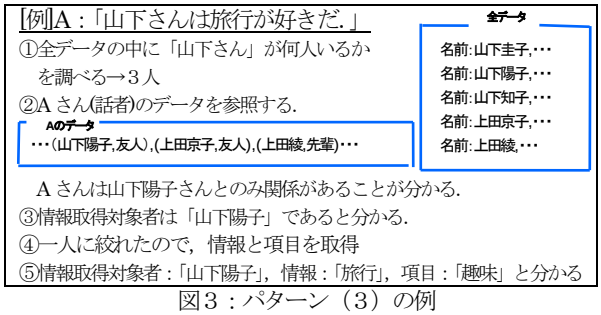

図３：パターン（３）の例

５．４　個人データ・項目取得方法 ～パターン(４)～

直前の文の主語が固有名詞であった場合，情報取得対象者は前文と同じ「○○さん」となる．例えば，A さんが，「山本さんは優しい人だ．」と言った後に，「彼女は料理がうまい」と言ったとすると，ここでの「彼女」は「山本さん」のことを指している．つまり，この様な場合は，対象者が特定できるので，次に情報・項目の取得を行う．一方，直前の文が無かったり，固有名詞以外であった場合は，誰のことを言っているのか特定できない．そこで，「彼(彼女)とは誰ですか？」と聞き返す．

５．５　個人データ・項目取得方法 ～パターン(５)～

初めに，話者と主語の人の人間関係を利用し，話者のデータの中から，主語の人の名前Aを探す．名前Aが見つかった場合，情報取得対象者はA さんとなり，情報・項目の取得を行う．

一方，名前A が見つからなかった場合，情報取得者の名前は分からないが，話者との関係は分かる．そこで，このような場合は話者との関係・情報・項目の３つを取得する．また，情報取得対象者が複数になる時は，情報・項目を取得するのと同時に，対象者の名前を提示し，「どの人のことですか？」と聞く．情報・項目の取得方法はすべて，主語が一人称の場合と同じである．

以上が，修飾語が一人称で主語が「人」の場合の処理の流れになっている．具体例を図４に示す．

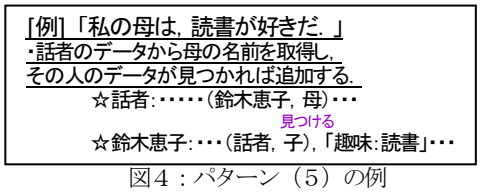

図４：パターン（５）の例

６．人間関係の表現方法

人間関係にはいろいろなものがあるが，本研究では，最も身近である血縁関係，サークル・会社関係，友人関係の３つについて表現方法を提案する．

まず，それぞれの関係にパターン番号をつける．血縁関係を１，会社・サークル関係を２，友人関係を３とする．

次に，座標を使って人間関係を表すことを考える．（パターン番号,ｘ座標,ｙ座標,性別番号)という形で記憶しておく．自分を原点と考え，ｘ座標で同世代の近さを表し，ｙ座標で地位の

上下を表している．また，性別番号とは，男女を区別するためのものであり，男性を１，女性を０とする．

６．１　血縁関係

会話に出てくる血縁関係の人物は，六親等内の人であることが多い．そのため，本研究でも，血縁関係として六親等内（又従兄弟まで）を対象とする．

自分を原点にした時の血縁関係を表したものを図５に示す．

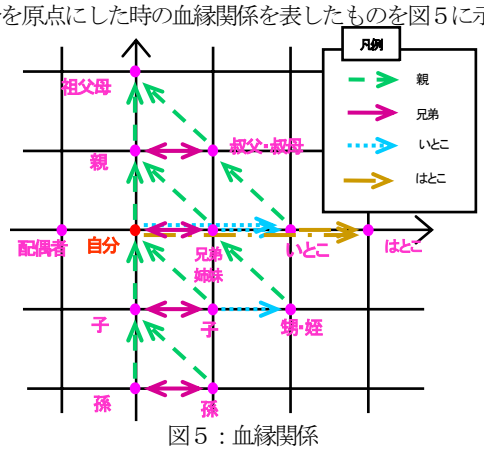

図５：血縁関係

ｘ＝０の時，ｙ軸正の方向を親，負の方向を子とすることで，親・子・祖父母・孫の位置が決まる．次に，ｘ軸正の方向に一目盛り進んだところを兄弟姉妹とすることで，叔父・叔母の位置も決まる．一方，ｘ≠０の時は，左斜め上の方向を親，右斜め下の方向を子とすることで，甥・姪，従兄弟の位置が決まる．また，これにより，ｘ＝０の時ｘ軸正の方向に二目盛り進んだところが従兄弟になり，それを利用することで，はとこが決まる．

６．２　会社・サークル関係

会社・サークルの関係も同様の方法で表すと，図６のようになる．

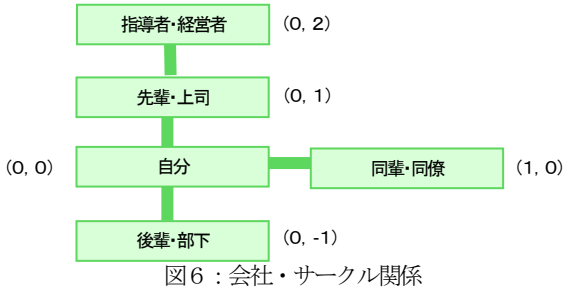

図６：会社・サークル関係

実際には，もっと多くの肩書きがあるが，自分との上下関係が分かればいいので，図６のようにする．

６．３　友人関係

友人関係を図７に示す．

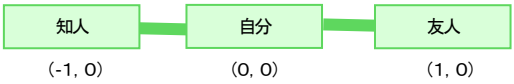

図７：友人関係

年齢などによる上下関係は考慮しないものとする．

７．実験評価

入力文は個性や個人の好みに関する文を中学校の英語の教科書などを参考に 200 文作成した．評価は，①情報を取得する人物が正しく認識できているか，②入力文から得られた情報とそれが入る項目の取得が正しくできているかの二点について目視により行う．

７．１　評価

７．１．１　人物取得に関する評価

人物の取得に関して評価を行った．一人の人物が正しく取得できた場合，または，人物が一人に特定できない時に正しい応答ができた場合を正解とした結果，78%の精度となった．５つのパターンに当てはまらないものを除いた 178 文についても同様の評価を行った結果，85%の精度となった．

７．１．２　情報・項目取得に関する評価

情報と項目の取得に関して，○・△・×の三段階で評価を行った．情報・項目とも正しく取得できている場合を○，正しい項目が取得できているが間違っているものも含まれている場合を△，正しいものが一つもない場合を×とした．また，人物取得の評価と同様，178 文についても評価を行った．200 文では，○が 33%，△が 39%，×が 28%となり，178 文では，○が 35%，△が 38%，×が 27%となった．○と△を正解とすると 200 文で 72%，178 文では 73%の精度であった．200 文の場合と 178 文の場合でほとんど精度に差が見られなかったことから，項目取得には，５つのパターンに当てはまるかどうかはそれほど関係していないと考えられる．また，△が多いため項目の選定条件が全体的に甘いと思われる．

７．２　考察・失敗例

７．２．１　人物取得に関する考察

7.1.1 の通り，200 文より 178 文の方が７%精度が良かった．これは，５つのパターンに当てはまらないもの(主語が２人，固有名詞＋「物」に関する名詞など)での失敗が多かったからである．５つのパターンに当てはまるにもかかわらず失敗した例としては，次のようなものがあった．

①表記一致による失敗

例：「私の叔父さんは公務員だ」

人間関係には「叔父」と登録してあったため，「叔父さん」が認識できなかった．このように，登録単語と違う表現をしたために不正解になっているものが他にも見られた．これについては，表記が違っていても，同じことであるということを理解させる必要がある．

②情報取得対象者が一人に絞れなかった場合

例：「彼女は犬を飼っている」

「私の祖母は犬が好きだ」の次に入力したので，本来なら祖母のことだと認識すべきなのだが，人間関係を調べた所，祖母が二人見つかり情報取得対象者が一人に絞れなかったため，「彼女とは誰ですか？」と聞き返してしまった．このように，情報取得対象者が複数の場合は，直前の入力で出てきた人物を記憶しておき，情報取得対象者として表示する必要がある．

７．２．２　情報・項目取得に関する考察

7.1.2 の通り，200 文の場合と 178 文の場合でほとんど精度に差が見られなかった．このことから，項目取得には，５つのパターンに当てはまるかどうかはそれほど関係していないと考えられる．また，△の所が多いため項目の選定条件が全体的に甘いと思われる．選定条件の再考が必要である．失敗した例としては次のようなものがあった．

①適切な項目がない場合

例：「私は夏が苦手だ」

普通なら，「夏：嫌いな季節」と考えるが，項目の中に「嫌いな季節」がなかったため，「夏：生年月日，年齢」となってしまった．このように，好みに関する適切な項目がなかったために失敗した例がいくつか見られた．

②性格と容姿の区別

例：「彼女はかわいい」

「かわいい：性格」となってしまった．これは，「用言」フレームの形容詞を「性格」とした所に問題があるのだが，これに関しては，「性格」と「容姿」を区別するルールを設定する必要がある．

③項目条件により削除された場合

例：「青木さんは京都生まれだ」

人間なら，「京都：青木さんの出身地」と分かるが，「京都生まれ」と「出身地」を結びつけることができなかった．「京都生まれ」と関連度が最も高いものは「出身地」になっているのだが，「出身地」の条件が都道府県名のみとなっているため，「京都生まれ」が削除されてしまった．

８．精度向上

評価結果を踏まえて，精度向上を行った．

８．１　表記変換知識ベースの導入

人間関係を参考にして人物取得を行う際，登録語と表記が一致しなければ認識できないという問題があった．そこで，表記のゆれをなくすために表記変換知識ベースを作成した．これは，登録語と違う表現をした場合，その語を登録語に変換するものである．一例を図８に示す．

・お父さん→父	・お兄ちゃん→兄
・おとうさん→父	・おにいちゃん→兄
・父さん→父	・兄ちゃん→兄
・父親→父	・兄貴→兄
・お母さん→母	・お姉さん→姉
・おかあさん→母	・おねえさん→姉
・母さん→母	・姉さん→姉
・母親→母	・お姉ちゃん→姉
・お兄さん→兄	・おねえちゃん→姉
・おにいさん→兄	・姉ちゃん→姉
・兄さん→兄	・姉貴→姉

図８：表記変換知識ベースの例

８．２　項目条件の再考

選定処理が甘いと言う問題があったため，項目の条件を再考した．まず，不足していた項目８項目を新たに付け加えた．また，好みに関する語（「好きな～」「嫌いな～」）に条件がほとんどついていなかったため，これらの項目に対して，シソーラスなどを使って条件を追加した．

８．３　再評価

８．３．１　人物取得に関する再評価

200 文で 92%，178 文で 96%の正解率になり，改良前と比べて，200 文では 14%，178 文では 11%精度が上がった．

８．３．２　情報・項目取得に関する再評価

○と△を正解とすると，200 文では 71%，178 文では 73%の精度であった．改良前と比べて，200 文・178 文ともほとんど精度に変化が見られなかったが，○のみの精度をみると，200 文の場合７%の向上が見られ，178 文の場合も８%の向上が見られた．

８．４　考察・失敗例

８．４．１　人物取得再評価の考察

人物取得評価については，改良を行ったことにより，精度はかなり上昇した．５つのパターンに当てはまるにもかかわらず失敗したものは，次のようなものである．

①情報取得人物の変化による間違い

例：「あいつはサッカー選手だ」

「私のおばあちゃんは和菓子が好きだ」の直後に入力したので，「あいつ」＝「おばあちゃん」と認識してしまった．確かに，「あいつ」が主語なので，これはパターン(４)に当たり，直前の文で情報を取得する人が特定できたらその人のことを続けて言っていると考えられる場合が多いのだが，おばあちゃんのことを「あいつ」ということは，まず考えられない．人間なら，ここで話題の人物が違う人のことになったと分かり，「あいつとは誰ですか？」と聞くのだが，その区別が現在のところではできていないために起こった間違いであると言える．

②名前入力による間違い

例：「英二君は生意気だ」

データには「名前：太田英二」と入っているのだが，名字と名前の区別ができず，性が「英二」となる人を探してしまったと考えられる．このように，名字ではなく下の名前を入力した時には失敗していた．

８．４．２　情報・項目取得再評価に関する考察

情報・項目取得については，○と△をあわせた精度は，改良前とほとんど変化が無く，○の精度のみが上がった．これは，項目条件を厳しくしたことで，改良前は不適切な候補が含まれて△になっていたものが，改良後に１つに絞られ○に変わったということを表している．また，何も拡張を行っていない場合つまり，取得した情報と最も関連度が高い項目を選んだだけの場合の精度も比較してみたところ，シソーラスを使って拡張したことで精度が約 30%向上していた．このことから，シソーラスを使って拡張し，項目条件を設定して選定処理を行う方法は有効であると考えられる．

９．おわりに

本研究では，入力文から話している相手の特徴や好みなど，個性に関する項目に当てはまる情報を取得する話者理解の方法を提案した．また，人間関係を理解させる方法も提案した．シソーラスを使って拡張を行い，各項目に条件をつけて選定処理を行うことで，入力文の中から必要な情報とそれが入る項目を取得することができた．項目取得方法や選定処理の方法を改良して精度を上げ，応答文生成に結びつけることが今後の課題となる．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト「知能情報科学とその応用」における研究の一環として行った．

参考文献

[1] 篠原宜道，渡部広一，河岡 司，“常識判断に基づく会話意味理解方式”，言語処理学会第８回年次大会発表論文集，B6-2，pp.651-654，(2002)

[2] 奈良先端科学技術大学院大学情報科学研究所 http://chasen.naist.jp/hiki/ChaSen/

[3] NTT コミュニケーション科学研究所　日本語語彙体系，岩波書店（1997）

[4] 渡部広一，河岡司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol8，No2，pp.39-54，(2001)

[5] 眞鍋康人，小島一秀，渡部広一，河岡司：“概念間の関連度やシソーラスを用いた概念ベースの自動精錬法”，同志社大学理工学研究報告，Vol.42，No.1，pp.9-20，(2001)