概念連想を用いた会話テーマ推定方式

奥田浩二　渡部広一　河岡司
同志社大学大学院工学研究科

1. はじめに

コンピュータを人間にとってより使いやすいものにするためには，人間同士が日常行っているコミュニケーション（意思疎通）の仕組みをモデル化し，コンピュータと人間とのインタフェースにこれを取り入れることが望まれる．その基本要素技術として，情報の意味を理解するメカニズムの研究が近年盛んに行われている．

そもそも人間どうしのコミュニケーションは，自分の伝えたい事柄（意図）を言葉にし，発話することで相手に伝えることと，相手の言葉を聞き取り，その意味を理解することの繰り返しによって成り立っている．そして，人は日々の経験を通じて今此処で行われている会話のテーマを適切に理解し，その趣旨に沿って話を展開していくことができる．しかし，会話内容を理解するための機能を持たないコンピュータが人のように適切な会話を行うことはできない．

そこで本稿では，語とその意味をセットにしたものをデータベース化した概念ベースを用いて，概念連想システムを構築し，それを利用して会話文中の単語からテーマを推定する手法を提案する．

2. テーマ推定

スポーツについて，料理について，旅行について... 人は自分が話していることのテーマを常に意識して会話を進めており，また会話相手も同様に共通のテーマを意識して会話を行っている．そして，そのテーマ周辺の知識を用いて会話を展開・深化させていくことで，会話を円滑に進めていく．たとえば，「A：夕食一緒に食べようよ．B：そうだね，じゃあ何料理を食べる？僕は和食がいいな．A：駅前に美味しいお寿司屋が出来たみたいだけど，そこに行ってみない？」という会話では，「食事」というテーマについて，「和食」，「寿司屋」という知識が用いられ，会話が展開されている．

本研究の目的は，会話文中からテーマに関連したキーワードを抜き出し，そこから会話のテーマを推定することである（図 1）．

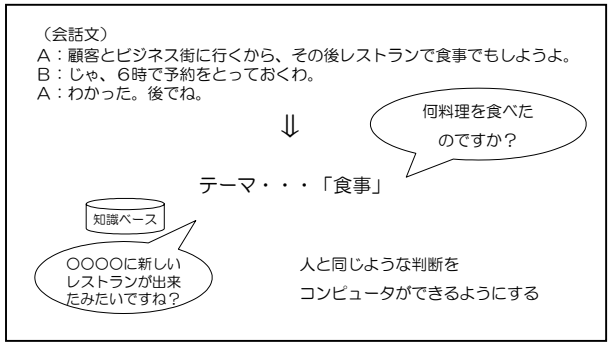

図 1：テーマ推定

3. 概念連想システム

提案する手法では主に N 語連想，逆引き N 語連想という概念連想システムを構築し用いる．以下に各連想システムとその要素技術について説明する．

3.1 概念ベース

ある語 A をその語と関連が強いと考えられる語 a_i と重み w_i の対の集合として定義する．

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_m, w_m)\}$$

a_i を一次属性と呼ぶ．また便宜上，A を概念表記と呼ぶ．このような属性の定義された語（概念）を大量に集めたものを概念ベース [1] と呼ぶ．

3.2 関連度

関連度 [2] とは，関連の種類しか特定できない意味ネットワークのようなものとは違い，概念と概念の関連の強さを定量的に評価するものである．関連度は，概念間の関連の強さを 0 と 1 の間の数値で表す．

3.3 シソーラス

シソーラス [3] には，一般名詞の意味的用法を表したものと用言の文型パターンを示したものがある．前者は，一般名詞の意味的用法を表す 2700 の意味属性（ノード）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り，約 13 万語（リーフ）が登録されている．また後者

は，日本語の用言 6000 語に対し，その用言がとる文型パターンを示したものである．

3.4　N語連想

N語連想とは，入力されたN個の単語から，そのすべての単語と最も関連が強い１つの単語を連想することである．具体的な方法は，概念ベースを用いて入力されたN個の単語の一次属性を順引きで展開し，そこから共通の属性を抜き出しその語を出力とする（図 2a）．共通の属性語が複数個あった場合は，関連度を用いて展開もとの単語との関連度を調べ，それを重みとしてその合計の最も高いものを出力として採用する（図 2b）．

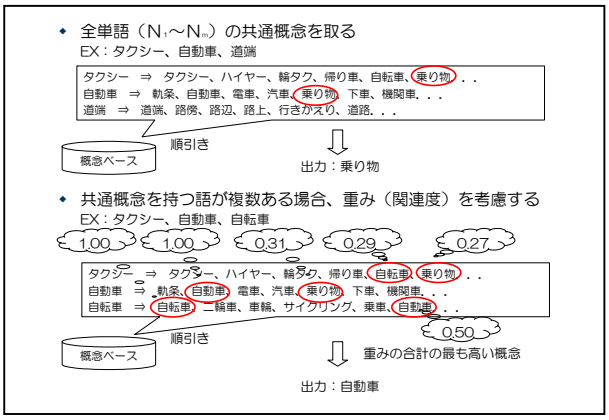

図 2：N語連想

3.5　逆引きN語連想

逆引き N 語連想の説明をする前にその要素技術である逆引き概念ベースについて触れておく．

逆引き概念ベースとは，概念ベースとは逆にある概念Xをその属性として持っているような概念群を，概念X属性として集めたものをデータベース化したものである．例えば「林檎」という概念をこの逆引き概念ベースで参照すると，「林檎」という概念をその属性として持っている，「梨」「蜜柑」「果物」などの語が呼び出される（図 3）．

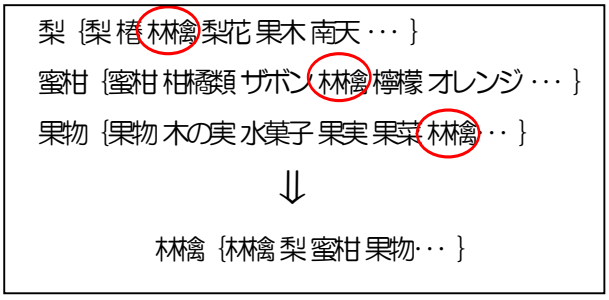

図 3：逆引き概念ベース

この逆引き概念ベースを利用して，N 語連想を行う手法が逆引き N 語連想である（図 4）．

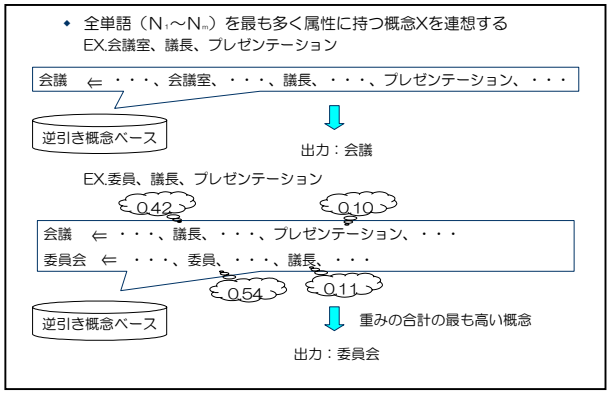

図 4：逆引きN語連想

4.　テーマ推定の手法

まず，会話文の種類を，意味的に近い単語が並んでいるタイプ（収束型）と，そうでないタイプ（発散型）の２パターンに分類する．例えば，「A：リー、今日の新聞のトップニュースは何？B：知りたいとは思わないだろう。とてもうんざりするものさ。A：大丈夫。うんざりするニュースには慣れてるから。」という会話文は前者にあたり，「A：すみませんがこの部屋では携帯電話を使ってはいけないことになっています。B：どうもすみません。知りませんでした。それでは外に出ます。A：ご協力ありがとうございます。ロビーのソファーを使用できますよ。」という会話文は後者に当たる．このように，会話中にある単語間の関連の強弱によって，会話文を収束型，発散型の２種類に分類する．

次に，N語連想，逆引きN語連想の特徴を見てみる．N語連想の特徴として，連想精度は良いが，出力が出ない場合があるという点が上げられる．一方，逆引きN語連想については，連想精度は低いが必ず出力が出るという点が挙げられる．この２点と，上記の会話文の２分類の性質を考えると，収束型の会話文に対してはN語連想を，発散型の会話文に対しては逆引きN語連想を適用すれば良いと考えられる．

また，収束型の会話文ではできるだけ関連の高い単語どうしをN語連想の入力とし，発散型では出来るだけ多くの単語を逆引きN語連想の入力とするべきであるという点から，それぞれ主語・述語，動詞・名詞を入力とするべきだろう．以上をもとに，図 5 のような処理の流れを考案する．以下，詳細に述べる．

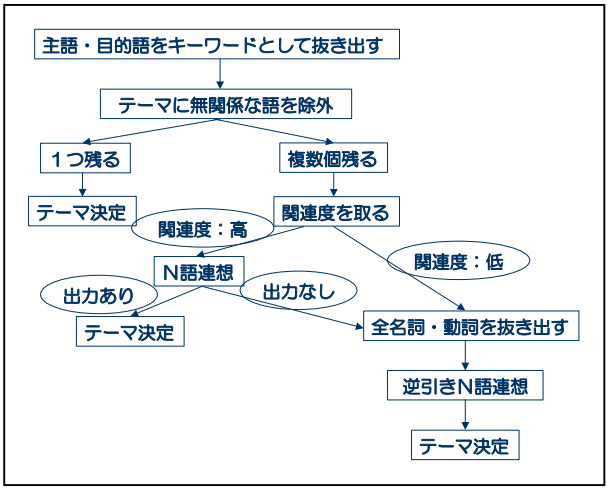

図 5：テーマ推定の手法

4.1　キーワードの選出

文中の重要な単語は主語・述語の２つであると考えられる．この２つをテーマ推定のためのキーワードとして活用する．

形態素解析を用いて，「名詞＋助詞“は，が，を，に”」という形になる名詞を主語・述語として抜き出す．次に，この中からテーマに関係のない語として，「固有名詞」，「代名詞」，「時間語」，「位置語」を除外する．「固有名詞」，「代名詞」の判別のために形態素解析を，「時間語」「位置語」の判別のために，それぞれシソーラスのノード「時間」，「位置，方向，方角」を利用する．そして，キーワードが１つ残った場合，そのキーワードをそのままテーマとして出力する．複数残った場合は次のステップに進む．

4.2　会話文の種類の特定

次に，会話文が収束型あるか，発散型かの判定を行うために，キーワードどうしの関連度を取る．この関連度の値が全体的に高ければＮ語連想に進み，低ければ逆引きＮ語連想に進む．判定方法は以下のようにする．

（１）他キーワードとの関連度が“0.04”以上のものが“Ｎ／2”個以上あるのとき，そのキーワードを高関連キーワードとする．

（２）“高関連キーワード数≧全キーワード数／2”ならＮ語連想に進む．
“高関連キーワード数＜全キーワード数／2”なら逆引きＮ語連想へ進む．

（３）Ｎ語連想に進む場合，低関連キーワードはキーワード群から除去する．

4.3　Ｎ語連想によるテーマ推定

会話文が収束型であると判別された場合，全高関連キーワードをＮ語連想にかけ，その出力をテーマとする．もし，出力がない場合は，逆引きＮ語連想に進む．

4.4　逆引きＮ語連想によるテーマ推定

逆引きＮ語連想の精度が低いという特徴を考えると，出来るだけ多くの情報を入力としてを与える必要がある．そこで，会話文が発散型と判別された場合，あるいはＮ語連想で出力がなかった場合，これまで保持していたキーワードを一度消去し，会話文中の全名詞・動詞をキーワードとして再設定し，テーマ推定のための情報量を増やしたうえで逆引きＮ語連想を行う．逆引きＮ語連想は出力が必ずあるので，出力結果をテーマとする．

5.　評価

5.1　全体の評価

テストデータとして，TOEIC の参考書[4) 5) 6)]から会話文を 100 セット収集し，このテストデータに対して，このシステムの推定結果がテーマとして妥当かどうか，人目で○：常識的，△：非常識でない，×：非常識の三段階で評価したところ，○：62%，△：16%，×：22%，○，△あわせて 78%の評価結果が得られた（図 6）．

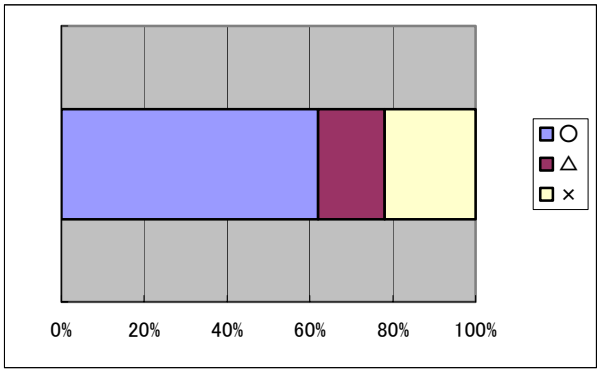

図 6：全体評価

5.2　Ｎ語連想，逆引きＮ語連想の評価

次に，キーワードがそのままテーマとなった 19 文，Ｎ語連想でテーマが決定された 31 文，逆引きＮ語連想でテーマが決定された 50 文の各評価を見る．キーワードがそのままテーマとなった場合，○：74%，△：10%，×：16%，○，△あわせて 84%，Ｎ語連想でテーマが決定された場合，○：65%，△：19%，×：16%，○，△

あわせて 84%，逆引きN語連想でテーマが決定された場合，○：56%，△：16%，×：28%，○，△あわせて 72%であった．

この結果から，N語連想の精度が逆引きN語連想の精度より高いことがわかる（図 7）．

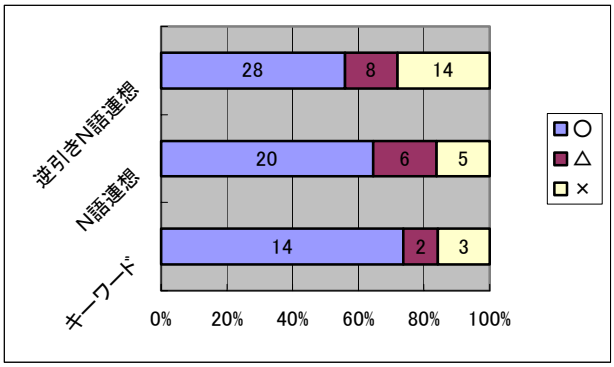

図 7：N語連想，逆引きN語連想の評価

5.3　正答例，誤答例

図 8 に正答例を掲載する．1 つめの例では，会話文中のキーワード「ＣＤ」がそのままテーマとして正しく出力されている．また，２つめの例では，キーワード「財布」「小銭」からN語連想により「お金」というテーマが正しく推定されている（図 8）．

◆ 正答例

（出力テーマ：CD　キーワード：CD）
A：CDをそんなに大きくかける必要はないよね．
B：ごめんなさい，うるさかった？
A：正直言うとね，明日朝が早いんだ．眠ろうとしてたけど眠れないよ．

（出力テーマ：お金　キーワード：財布,小銭）
A：２０ドル崩せるかい？
B：分からないわ．財布を見てみる。
A：ありがとう．自動販売機を使うのには小銭がいるんだよ．

図 8：正答例

次に，誤答例を掲載する．1 つめの例では，会話文中のキーワード「老人」がそのまま間違ったテーマとして出力されている．また，２つめの例では，キーワード「切符」「割引」からN語連想により間違った「お金」というテーマが推定されている．これは，キーワード「切符」「割引」を概念ベースで属性展開し，共通属性をとったさいに，「手形」といった間違った共通属性がテーマとして出力されてしまったからだと考えられる（図 9）．

◆ 誤答例

（出力テーマ：老人　キーワード：老人）
A：このコーヒー、ブルックリンの小さなお店で買ったの。
　面白い老人が経営しているの。
B：本当においしいね。そこへ行く道教えてくれない？
A：もちろん。それがお望みなら私が買ってくるよ。

（出力テーマ：手形　キーワード：切符,割引）
A：往復切符はいくらですか？
B：片道料金の２倍です。
A：割引は無いということですね。

切符｛入場券,券,乗車券,切符,食券,チケット,回数券,割符,債権,挨拶,
　い札,クーポン,手形,馬券,印紙,切手･･･｝
割引｛割引,手形,分引き,引き出物,割引債,割り増し,価格,引退,
　割り切る,引っ張る,差し引き,歩合,割合,何分,退却･･･｝

図 9：誤答例

6.　おわりに

本論文では，N語連想，逆引きN語連想という 2 種類の概念連想システムを構築し，それを適切に組み合わせ活用することで，会話文のテーマを適切に推定するモデルを提案した．会話文のテーマが明確になることで，コンピュータが会話を進めていく上での重要な手がかりが得られ，会話が円滑に進むことが期待される．

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における研究の一環として行ったものである．

参考文献

1) 広瀬幹規，渡部広一，河岡 司，“概念間ルールと属性としての出現頻度を考慮した概念ベースの自動精錬手法”，信学技報，NLC2001-93，pp.109-116，2002.
2) 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，Vol. 8，No. 2, pp. 39-54，2001.
3) NTT コミュニケーション科学研究所監修，「日本語語彙体系」，(岩波書店，東京，1997)
4) 浅見ベートーベン，“TOEIC TEST リスニング完全征服法”，日本 IBM（株），2001.
5) Park Deuk-Woo，Choi Byong-Gil，“TOEIC テスト 730 点攻略本”，旺文社，2000
6) 菊間ひろみ，“これだけ！TOEIC TEST 総合対策”，あさ出版，1999