言語の非線形性に着目した連想システムの構築

奥村 紀之
長野工業高等専門学校電子情報工学科
noriyuki_okumura@ei.nagano-nct.ac.jp

1 はじめに

本研究は，人間とコンピュータが円滑にコミュニケーションをとるために重要な基盤技術である連想システムとそれを支える概念ベースの応用研究である．近年，言語の非線形性が着目されており，池原の著書でその重要性が指摘されている [1]．文献では，置換可能な言語要素を線形要素，置換不可能な言語要素を非線形要素として定義し，非線形性に着目した言語処理手法が提案されている．本稿では，いずれのユーザに対しても適用可能な連想語，話題を線形要素，ユーザ個々の嗜好に応じて代替が不可であるような連想語を非線形要素としてとらえている．

我々はこれまでに，人間のように柔軟な連想機能をコンピュータで実現するために，概念ベースと関連度計算方式の開発を行ってきた，しかし，あらゆるユーザに対して画一的な回答を返すシステムとなっており，まだまだ柔軟性に乏しいのが現状である．

そこで，対話対象となるユーザの情報をコンピュータに付与し，ユーザに応じて会話の流れに即した話題を呈示するシステムを開発することによって，個々のユーザに合わせた対応が可能となるシステムを開発することを目指す．

本稿では，ユーザの情報としてニュース記事，コラム，教科書を代替要素とし，これらをコンピュータに付与することにより，ユーザとの対話の中で次に呈示すべき話題を抽出する手法について検討している．

2 関連技術

2.1 概念ベース

人間と機械が会話によって意思疎通を図るためには計算機に言葉の意味を理解させる必要がある．計算機に言葉に関する知識を付与するための知識ベースとして，単語の意味を定義している「概念ベース」と呼ばれる, 辞書に類似したデータベースがある [7]. 概念ベースにおいて, 任意の概念 A は, 概念の意味特徴を表す属性 a_i と属性 a_i が概念 A を表す上でどれだけ重要かを示す重み w_i の対の集合として定義している (式 1).

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\} \quad (1)$$

ここで,a_i を一次属性と呼び，A を概念表記と呼ぶ．このような (概念-属性) の集合を大量に集めたものを概念ベースと呼ぶ．ただし, 任意の一次属性は, その概念ベース中の概念表記の集合に含まれているものとする．すなわち, 属性を表す語もまた概念として定義されている．したがって, 一次属性は必ずある概念表記に一致するので, さらにその一次属性を抽出することができる. これを二次属性と呼ぶ.

また, 概念ベースを用いて, 概念間の関連の強さを推し量る関連度計算方式, 共起度計算方式がある．以下にこれらの計算方法について詳細を述べる．

2.2 関連度計算方式

関連度とは, 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．本研究では, 関連度計算方式に，意味関連度計算，意味的共起関連度計算 [6] の 2 種類を用いた．以下に関連度計算方式の詳細を示す．

2.2.1 意味関連度計算

概念の意味属性の一致度合いから関連度を評価する意味関連度計算では，意味的に近い単語に対して精度の良い関連度計算が行える．以下に, 意味属性を考慮した関連度計算方式である，重み比率付き関連度計算方式の詳細を示す.

任意の概念 A,B について, それぞれ一次属性を a_i,b_j とし, 対応する重みを u_i,v_j とする．また概念 A,B の属性数を L 個,M 個 $(L < M)$ とする．

$$A = \{(a_i, u_i) | i = 1 \sim L\}$$

$$B = \{(b_j, v_j)|j = 1\sim M\}$$

このとき, 概念 A,B の重み比率付き一致度 $MatchWR(A, B)$ を以下の式で定義する.

$$MatchWR(A, B) = \sum_{a_i=b_j} min(u_i, v_j)$$

$$min(\alpha, \beta) = \begin{cases} \alpha & (\beta > \alpha) \\ \beta & (\alpha > \beta) \end{cases}$$

このように一致度を定義するのは,$a_i = b_j$ となる属性に対し, 互いの属性の重みの共通部分が有意に一致すると考えるからである.

次に属性の少ない方の概念を A とし ($L \leq M$), 概念 A の属性を基準とする.

$$A = \{(a_1, u_1), (a_2, u_2), \cdots, (a_L, u_L)\}$$

そして概念 B の属性を, 概念 A の各属性との重み比率付き一致度 $MatchWR(a_i, b_{xi})$ の和が最大になるように並び替える.

$$B = \{(b_{x1}, v_{x1}), (b_{x2}, v_{x2}), \cdots, (b_{xL}, v_{xL})\}$$

これによって, 概念 A の一次属性と概念 B の一次属性の対応する組を決める. 対応にあふれた概念 B の属性は無視する (この時点では組み合わせは L 個). 但し, 一次属性同士が一致する (概念表記が同じ) ものがある場合 ($a_i = b_j$) は, 別扱いにする. これは概念ベースには 9 万の概念が存在し, 属性が一致することは稀であるという考えに基づく. 従って, 属性の一致の扱いを別にすることにより, 属性が一致した場合を大きく評価する. 具体的には, 対応する属性の重み u_i,v_j の大きさを重みの小さい方にそろえる. このとき, 重みの大きい方はその値から小さい方の重みを引き, もう一度, 他の属性と対応をとる. 例えば,$a_i = b_j$ で $u_i = v_j + \alpha$ とするば, 対応が決定するのは (a_i,v_j) と (b_j,v_j) であり,(a_u,α) はもう一度他の属性と対応させる. このように対応を決めて, 対応の取れた属性の組み合わせ数を T 個とする.

重み比率付き関連度とは, 重み比率付き一致度を比較する概念の各属性間で算出し, その和の最大値を求めることで計算する. これを以下の数式により定義する.

$$Rel(A, B) = \sum_{i=1}^{T} MatchWR(a_i, b_{xi}) \times (u_i + v_{xi}) \times (min(u_i, v_{xi})/max(u_i, u_{xi}))/2 \quad (2)$$

以下, 重み比率付き関連度を関連度と略す. 関連度の値は 0〜1 の連続値をとり,1 に近づくほど概念間の関連性が高い.

2.2.2 共起度計算方式

共起度とは, 概念と概念が同時に出現する (共起する) 度合いを示したものである.

まず, 任意の表記 a,b を属性として持つ概念を A_i,B_j とする. また, 表記 a,b を属性として持つ概念の個数を L 個,M 個とする.

$$a = \{A_i|i = 1\sim L\}$$
$$b = \{B_j|j = 1\sim M\}$$

このとき,$A_i = B_j$ となる (共起した) 概念の個数を $a \cap b$ で定義する.(図 1)

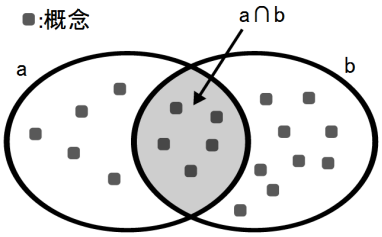

図 1: $a \cap b$ **のイメージ図**

$a \cap b$ は表記 a,b の共通の値となるが, 共起の割合は表記 a と表記 b において異なる. そのため共起度は, 表記 a,b の共起の割合の平均とする. 数式は以下のように定義する.

$$Co(a, b) = (\frac{a \cap b}{L} + \frac{a \cap b}{M})/2 \quad (3)$$

共起度の値は 0〜1 の連続値をとり,1 に近づくほど共起している度合いが強い.

2.2.3 意味的共起度関連度計算方式

先述した意味関連度計算は, 意味的に近い単語に対して精度の良い関連度計算が行えたし. それに対し, 表記的共起関連度計算方式は, 意味関連度計算に加え, 概念の共起情報を用いることで, 人間が連想によって導き出せる単語間の関連度の評価を可能にした [6].

しかし，関連が強いと思われる概念同士でも，その概念表記が全く共起しない場合も存在する．そこで，奥村らは，概念の一次属性だけでなく，二次，三次と市世湯する属性の範囲を広げることで，関連の強い疑念同士では共起率が高くなる．概念 A の n 次属性に概念表記 B が多数現れ，概念 B の n 次属性に概念表記 A が多数現れるという性質から，奥村らは，概念の関連性を用いて関連度を判断する方法を，意味的共起関連度計算方式として定義している [6].

2.3 シソーラス

シソーラスは，単語を意味的に分類した分類体系である [4]．シソーラスの多くは木構造を持ち，名詞の集合を分類した名詞シソーラスや，用言の集合を分類した用言シソーラスなどがある．また，木構造の葉（以下，リーフと呼ぶ）のみに単語が所属する分類シソーラスと，根及び中間ノードにも単語が所属する上位下位シソーラスがある．本研究では，木構造を持つ名詞シソーラスであり，上位下位シソーラスの 1 つである NTT シソーラスを用いる．NTT シソーラスは一般名詞の意味的用法を表す 2710 個のノードの上位－下位関係，全体－部分関係が木構造で示されたものである．ノードに所属する名詞として約 13 万語のリーフが分類されている．図 2 に NTT シソーラスの木構造の一部を示す．

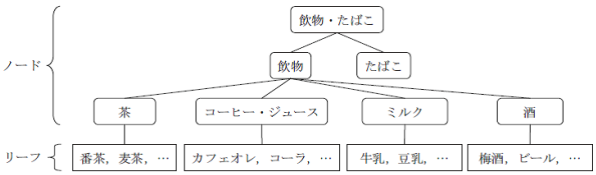

図 2: **シソーラスの構成図**

また, 一般に木構造で上位に属する概念をノード, 下位に属する概念をリーフと称する．本研究では，注目する単語の上位に属するノードを親，下位に属するリーフを子, 同じノードを親に持つリーフを兄弟と定義する．親，子，兄弟の関係図の例を図 3 に示す．

また，ここで更に本研究内で使用する連想の概念について定義する．図 3 において，注目している単語から見て，親と子のように，木構造上で一段階離れている単語を，”一段階連想可能な単語” と定義する．また，兄弟のように，同じノードを親に持つ単語を，”二段階連想可能な単語” と定義する．注目する単語の親, 子,

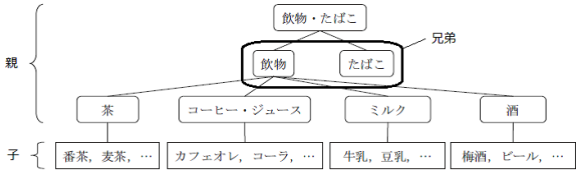

図 3: **シソーラスの親子関係図**

兄弟よりも関係が離れている単語に関しては，連想不可能な単語とする．

3 提案手法

本システムは，ユーザーから話題の入力を受け取り，話題から主題語と話題語を取得する．その語を元に，背景的知識から話題となりうる語を抽出し，話題候補語群とする．更に，主題語と話題語，話題候補群を用いて話題の適合性を判断する．図 4 に本研究のシステムの構成図を示す．

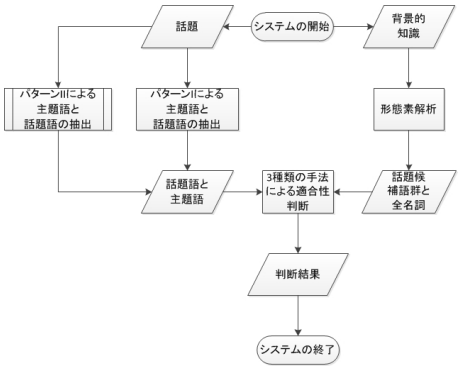

図 4: **システムの構成図**

3.1 背景的知識

林は，会話の背景に存在する膨大な知識を「背景的知識」と呼称している [2]．本研究では，ある文章を一定量の知識と見立て，人間の知識の代わりに利用した．背景的知識に利用した文章の種類は, ニュース，コラム，教科書の 3 種類である．

3.2 主題語・話題語

野田らはある主題語 p の話題となる語 t は,「p の t」の形で用いる傾向にあると述べている [3]．よって, 本稿での話題は,「主題語の話題語」という形式，もしく

は日常会話でよく使用される「形容詞＋名詞」の形で表されるものとする.

3.3 話題候補語群

背景的知識内の主題語が出現している文章内から，名詞・形容詞を抽出し，これを話題候補語群とする. これは，話題は「主題語の話題語」の形で表すことができるため，主題語が出現している文章内の単語は，背景的知識内の他単語に比べて話題になり得る可能性が高いと考えるためである.

4 評価実験

本研究では，話題の適合性を判断する手法，適合性判断方法を提案する. また，適合度計算の他に，主題語と話題語の決定方法を提案する.

5 評価方法

システムの評価実験は，3 種類の背景的知識から作成した，以下の 4 種類の話題で行う.

1. 文章中の単語を含めた、文章に確実に関係がある話題
2. 文章中の単語から連想できる単語を含めた、文章に関係がある話題
3. 文章中の単語から連想した単語から、更に連想できる単語を含めた文章には関係がない話題
4. 文章に全く関係がない話題

この 4 種類の評価データにおいて，1，2 の話題は背景的知識に対する話題としてふさわしい話題，3，4 の話題は背景的知識に対する話題としてふさわしくない話題とする. 前者を適合，後者を不適合として評価実験を行う.

5.1 主題語と話題語の決定方法

話題を形成する単語が 3 つ以上である場合の主題語と話題語の決定方法について提案する. 評価データから名詞と形容詞を取得し, 取得した順に A,B とアルファベットを割り当てる.

パターン I

パターン I では，B を主題語，C を話題語とする.

パターン II

パターン II では，それぞれ A-B，B-C，A-C の共起度を計算し, 結果値から動的に主題語，話題語を決定する.

適合性判断手法 A

3.1 節で示した評価用データの作成条件に基づいて適合性の判断基準を設定した手法である.

適合性判断手法 B

適合性判断手法 A の，類似語を含む話題を不適合と判断する問題点を考慮した判断手法である.

A では，背景的知識内の主題語，話題語の出現非出現を重要視しているため，主題語と話題語に関連する単語が出現していない話題は不適合と判断した. しかし，不適合と判断された話題には，以下のような傾向が見られた.

- X:主題語が背景的知識内に出現，話題語が非出現かつ，背景的知識内の名詞に話題語の兄弟 (X') が含まれる.
- Y:主題語が背景的知識内に非出現，話題語が出現かつ，背景的知識内の名詞に主題語の兄弟 (Y') が含まれる.

X，Y に該当する話題における X'，Y' は，主題語と話題語の類似語である傾向にあった. ここで，兄弟語はシソーラスにおいて，同じノードを親に持つリーフ同士を指す.

以上の問題を考慮し，適合性判断手法 B では兄弟語を使用した. 単に兄弟語が存在すれば適合と判断すると，適合判断の許容範囲を広げるだけになる. X では話題語と X' の，Y では主題語と Y' の関連度計算を行い，関連度がしきい値以上だった場合，適合と判断した.

表 1: 各適合性判断手法における正答率表

適合性判断手法	正答率 (%)
IA	68.9
IIA	66.6
IB	71.8
IIB	68.2

表 2: 適合性判断手法 B における類似語の判定例

判定	主題語	話題語	A	B
成功	母親	団体	不適合	適合
失敗	母親	集まり	不適合	不適合

5.2 結果

パターン I で主題語と話題語を選出し，適合性判断手法 A を用いた判断手法を IA，他も同様に，IB，IIA，IIB とする. 表 1 に各手法の正答率を示す.

ただし，ここに示すデータはコラム [5] から抜粋した文章を使用している.

図 5 に，評価データの種類別の正答率を示す.

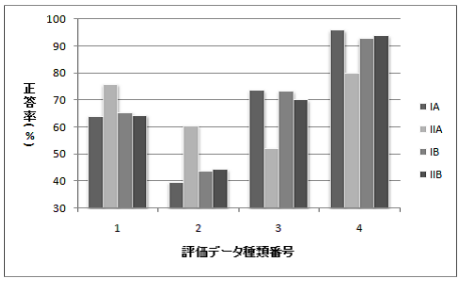

図 5: 各適合性判断手法における話題種類別正答率

6 考察

パターン II では主題語と話題語の選択を動的に行うことで，I に比べ正確な話題を選択できた. 正しい話題を使用し, 関連語を多く抽出できたことで, 適合の許容範囲が広くなったため，不適合のデータに対する正答率が低くなった結果, パターン I に比べ全体の正答率が低い. しかし, ユーザーの話題が 2 語で構成されるとは限らないため，適合性判断において主題語と話題語の動的決定は重要である. 以上から，パターン II の精度を向上させることで，適切な適合性判断が行えると考える.

適合性判断手法 B では，表 2 のように,A では適正な判断ができなかった話題対して正しい判断を行えた. また, 表 1 から,IA の正答率 68.9% に比べ，IB は 71.8%と約 3%の差異が確認できた. 適合性判断手法 B は，A に比べ適切な適合判断が行える手法であるといえる.

7 まとめ

本稿では，背景的知識としてニュース，コラム，教科書を利用し，ユーザの情報としてコンピュータに付与し，話題を選出する手法を検討した. 特に，ユーザとの会話において継続的な話題を提供することは容易ではあるが，ユーザに応じて提供する話題を飛躍させるという処理の実現が，ユーザとの会話を発展させる上で非常に重要となるため，本稿で検討したような話題の適合性判断が有効に活用されるものと考える.

また，コンピュータの判断基準に非線形性を持たせることによって，ユーザに応じた柔軟な処理を実現するための手法を検討していく必要がある.

参考文献

[1] 「非線形言語モデルによる自然言語処理」池原悟, 岩波書店，2009 年

[2] 「相手の嗜好にあった話題を提供する自動発話システムの開発」林 輝大 : 第 72 回情報処理学会全国大会　6X-7 2010 年 3 月

[3] 「主題語からの話題語自動抽出とこれに基づく Web 情報検索」野田 武史，大島 裕明，小山 聡, 田島 敬史, 田中 克己 : 電子情報学会 , DE2006-90

[4] 「日本語語彙体系」池原 悟，宮崎 正弘，白井 諭, 横尾 照男，中岩 浩巳，小倉 健太郎，大山 芳史, 林 良彦：岩波書店. 1997

[5] 「ない不便は本当に不便か」大平　一枝 : http://www.asahi.com/housing/diary/TKY201011080114.html

[6] 「概念の意味属性と共起情報を用いた関連度計算方式」渡部 広一, 奥村 紀之, 河岡 司 : 同志社大学理工学研究報告 第 48 巻 第 3 号

[7] 「概念間の関連度計算のための大規模概念ベースの構築」奥村紀之，土屋誠司，渡部広一，河岡司: 自然言語処理，Vol.14，No.5，pp.41-64，2007.