語の階層構造とユーザーの嗜好情報に基づく話題選出手法

林 輝大†
†長野高専 専攻科電気情報システム専攻
10912@st.nagano-nct.ac.jp

奥村 紀之‡
‡長野高専 電子情報工学科
noriyuki_okumura@ei.nagano-nct.ac.jp

1　はじめに

近年の情報社会において自分の欲しい情報だけを抽出するのは難しい. そこで本研究では計算機と人間が会話を行うことでこの問題の解決を図る.

本論では, 会話における話題の選択に着目し, 2つの指標を提案している. 1つは Wikipedia のデータから語の階層構造の構築し, 得られた階層を指標とする手法. もう1つは Twitter のログから Twitter ユーザーの嗜好情報の抽出し, ユーザーと語の関係を指標とする手法である. そして, それらの指標と概念ベースを利用した話題選出手法を提案している. 実験ではその指標を基準に選出した話題の傾向とその精度を評価している.

2　提案手法

話題選択の指標は, 語の階層構造および嗜好情報から算出している. 手法の詳細を以下に記述する.

2.1　話題語の種類

本研究で会話における話題を示す語を「話題語」と呼称する. また, 次の話題の候補として挙げられる話題語の事を「候補語」と呼称し, 複数の候補語がある場合「候補語」して呼称する. さらに候補語は3種類にわけ, 以下のように定義する.

継続語 現在の話題語と本質が同じ話題語

発展語 現在の話題語に関連した話題語

新規語 現在の話題語と関係ない新しい話題

2.2　話題語同士の関係の定量化

無数に存在する話題語から特定の話題語を抽出するためには, 話題語同士の関係を定量化し, 適不適の判断を可能にする指標を設ける必要がある. そこで本研究では, 語の意味を定義している概念ベース [1] を用い, 話題語の関係を定量化して指標としている.

2.3　語の階層構造の構築

会話は現在の話題からより具体的, もしくはより抽象的な話題を選択される場合がある. そこで本研究では日本語版 Wikipedia のダンプデータから抽出した上位下位語より語の階層構造の構築を行い, その話題選択の指標として用いることを提案する.

2.3.1　日本語語彙体系との対応付け

構築した語の階層構造の特徴として, 上位語が0個の語が11万語近くある. これらは最上位の語であるが, 全てが同じ階層であると判断するのは困難である. そこで柴木ら [3] の手法を参考にして, 既に最上位語が1個に定められている日本語語彙体系を用い, 11万語と対応を取ることで, 階層の決定を行う.

対応付けの方法は以下の3段階で行う.

1. Wikipedia より得られた最上位語 (A) を形態素解析する.
2. 解析結果の中で一番最後に現れた名詞 (B) と一致する階層を日本語語彙体系より得る.
3. A の階層を B の階層より1つ下として扱う.

「うどん料理」の一例

STEP1：うどん/料理 (形態素解析)
STEP2：「料理」の階層→ 10(∵日本語語彙体系)
STEP3：「うどん料理」の階層→ 11

以上の方法で最上位語の 94.4%(111,158 / 117,745) が対応ができた. また対応付けした最上位語の中から 1,500 の標本を無作為に選出し, 上位下位関係が正しいか評価した所, 87.8%(1,317 / 1,500) の精度であることがわかったため, この手法で最上位語の階層を決定

し, それ以下の語の階層を決定する. これにより生成した語の階層構造は対応前と比較して 99%(2,918,177 / 2,946,961) の語に対応付けができた.

2.3.2 語の階層構造と話題語の関係

階層構造から得られる階層を話題選択の指標として用いる. ある話題語 A, B の階層が X, Y であるとしたとき, A と B の深度差を以下の式で定義する.

$$Depth(A, B) = |X - Y| \quad (1)$$

2.4 Twitter データからの嗜好情報の抽出

会話では会話相手の嗜好を推測しながら, 話題を選択する場合がある. そこで Twitter のログデータより発言の傾向を解析し, 嗜好情報の抽出を行い, 話題選択における指標として用いることを提案する.

2.4.1 Twitter のログデータの解析

解析対象の Twitter のログデータは 449 ユーザーの一年分 (2010 年 1 月〜2010 年 12 月) の発言を用いている. 解析は IBM が提供している「IBM Content Analytics」を使用している.

IBM Content Analytics により解析を行うと, 語の頻出度, 相関などが得られる. 本研究ではユーザーと一般名詞との相関の値を使用する. 一般名詞を話題語と捉え, 以降ユーザー u と話題語 w の相関値を v_{uw} と定義する.

2.4.2 嗜好情報と話題語の関係

Twitter データの解析結果より得られた相関値 v_{uw} を嗜好情報と捉え, 話題選択の指標を算出する.

話題語 a と相関の持つユーザー $u_i(i = 1〜l)$ 内で, 話題語 b と相関のあるユーザー数を n_{ab} と定義する. ただし, ユーザー u が話題語 b と相関があると判断するのは $v_{ua} \neq 0$ かつ $|v_{ua} - v_{ub}| \leq \alpha \ v_{ua}$ であるときとする. このとき α は許容値を示すものであり, $0 < \alpha < 1$ とする. 会話相手がある語 x に関係があることが既知の時, 話題として提供できると推測した候補語群 $R(x)$ を以下の式で定義する.

$$R(x) = \{(w_1, n_{xw_1}), (w_2, n_{xw_2}), \cdots, (w_m, n_{xw_m})\} \quad (2)$$

w_m は候補語であり, n_{aw_m} は次の話題語として推薦する度合いとなる.

しかし n_{ab} は値の範囲がユーザー数に左右されてしまうので, n_{ab} を拡張した指標として, 正規化された値 e_{ab} を提案する. e_{ab} を用いた時の候補語群 $ER(x)$ を以下のように定義する.

$$ER(x) = \{(w_1, e_{xw_1}), (w_2, e_{xw_2}), \cdots, (w_m, e_{xw_m})\} \quad (3)$$

3 評価実験

今回の実験では, 語の階層構造および嗜好情報を用いた指標が話題選択においてどのような傾向を示すか, またどの程度の精度を持つのかを評価する.

3.1 予備実験

式 (1) と候補語群との間に, どのような関係にあるかを調査した. 現在の話題語を三種類 (「勉強」「動物」「旅行」) 選択し, それぞれの候補語群との深度差を式 (1) により算出する. 得られた結果の平均を図 1 のようにグラフにした. 図 1 は縦軸が候補語の数, 横軸は深度差を示している.

図 1 を見ると, 深度差 0〜200, 深度差 200〜800, 深度差 800 以上に特徴が見られ, 3 つのグループに分けることが可能であると判明した.

このことから, 深度差 0〜200 を近傍グループ, 深度差 200〜800 を中間グループ, 深度差 800 以上を遠方グループと呼称する.

3.2 実験 1

予備実験でわけられた 3 つのグループの特性の調べ, 語の階層構造から得られる指標の性質を調査する.

3.2.1 実験方法

現在の話題語を三種類 (勉強, 動物, 旅行) 用意し, 概念ベースからそれぞれの二次属性までを獲得し, それを各候補語群とする. 候補語群を式 (1) により 3 つにわけ, 概念ベースを用いた関連度により降順に並び替える. 各グループの上位 10 個, 計 30 個の候補語を被験者 3 名に見せ, 継続語, 発展語, 全く関係ない話題語の内, どれに属するかを判断させた.

3.2.2 実験結果

現在の話題語に対して適切であると判断された話題語が候補語群に含まれる率を含有率と呼称し, 以下のように定義して精度の判断基準として用いる. ただし適切である話題語とは継続語, 発展語のいずれかを指す.

$$\text{含有率} = \frac{\text{適切と判断された候補語の数}}{\text{対象となる候補語の数}} \quad (4)$$

またもう 1 つの評価尺度として, 以下の式を用いる.

$$\text{再現率} = \frac{\text{該当数}}{\text{適切と判断された候補語の数}}$$

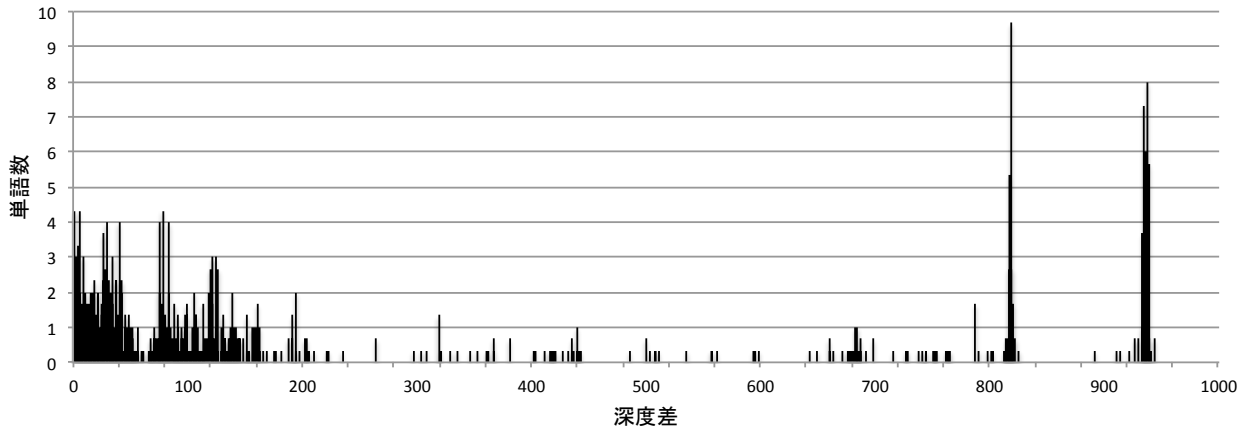

図 1: **階層深度と候補となる単語の数の関係**

$$\text{適合率} = \frac{\text{該当数}}{\text{対象となる候補語の数}}$$
$$F\text{ 値} = \frac{2\ \text{再現率}\ \ \text{適合率}}{\text{再現率} + \text{適合率}} \quad (5)$$

該当数とは, 継続語と判断された候補語の数の場合と, 発展語と判断された候補語の数の場合の 2 通りを考慮する.

以上, 2 つの評価尺度で実験結果をまとめたのが, 表 1, 2 である.

表 1: 各グループにおける含有率 (被験者 3 名)

	近傍	遠方	中間	全体
含有率	**92.2%**	70.0%	77.8%	80.0%

表 2: 各グループにおける F 値の平均

該当	近傍	遠方	中間
継続語	**0.566**	0.324	**0.464**
発展語	0.393	**0.495**	0.402

3.2.3 考察

まず表 1 から近傍グループの含有率に関しては 92.2%という数値が出ている. 先行研究 [2] においては 87%の値であったので, 比較的高い値ということがわかる.

続いて各グループの F 値の平均をみる. この F 値より候補語群に偏りがあるかどうかがわかる. 表 2 より, 近傍と中間のグループは継続語, 遠方のグループは発展語と F 値が比較的高く, それぞれに傾向があると予測できる. 実際に有意な差があるかどうかを F 検定および T 検定により判断を行った. その結果, 近傍と遠方の F 値の平均に優位な差が見られ, 近傍において継続語, 遠方において発展語が候補語としてを比較的多く含まれることが判断できる.

これらの結果から式 (1) から算出された深度差は, 候補語をグループによっては偏った選別をする性質があることがわかる. 従って, 継続語か発展語を選択する指標として用いることが可能であると推察する.

3.3 実験 2

この実験では, 嗜好情報から算出した式 (2), (3) で得られる候補語群 R, ER に対し, ユーザー数がどのような影響を及ぼすのか, その傾向調査を行う.

3.3.1 実験方法

ユーザー数を 5, 10, 50, 100 人と変化させた時, 想定する会話相手の情報を 3 つ (勉強, 授業, 学校) 設定し, 候補語群 R, ER を算出する. 獲得した候補語群 R, ER を n_{ab}, e_{ab} で降順に並び替え, 上位 20 個の候補語を算出する. 重複を除き 309 語の候補語を被験者 3 名に見せ, 話題として提供できるかどうかを判断させた.

3.3.2 実験結果

適切である候補語を, 被験者の過半数が話題として提供できると判断した語とし, 含有率を求める. ユーザー数別の含有率を表 3 にまとめる.

3.3.3 考察

ユーザー数の増加に伴い, 含有率も高くなる傾向が見られた. これはユーザー数が増加することで候補語の幅が広がるからであると推察する. また候補語群の

表 3: ユーザー数別の含有率 (被験者 3 名)

ユーザー数	5	10	50	100
R	**62.5%**	**60.0%**	73.3%	71.7%
ER	55.8%	55.0%	**76.7%**	**75.8%**

比較では, n_{ab} がユーザー数に依存すること, ユーザー数が多い場合において ER の方が含有率が良いことから, 候補語群 ER を使用するのが良いと判断する.

3.4 実験 3

この実験では, 実験 1,2 で得られた知見をもとに, 語の階層構造と趣向情報から得られた指標を組み合わせた時, 精度にどのような影響を及ぼすのか評価する.

3.4.1 実験方法

現在の話題 3 つ (動物, 旅行, 勉強) を用意し, 会話相手の情報を 3 つ (勉強, 授業, 学校) 設定する. 候補語群は, 嗜好情報をユーザー数 100 人で抽出した時の ER とする. 得られた候補語群を深度差により 3 つに分けた上で, 各グループの候補語を以下の式で降順に並び替え, 上位 10 個, 計 30 個を選出する.

$$\text{関連度} \quad (1 + e_{ab}) \tag{6}$$

選出した候補語を被験者 3 名に見せ, 継続語, 発展語, 新規語, 全く関係ない話題語の 4 つを判断させた.

3.4.2 実験結果

適切な候補語を、継続語、発展語、新規事して含有率を求める。各グループにおける含有率と全体の含有率を表 4 にまとめた.

表 4: 階層構造と嗜好情報を組み合わせた場合の含有率

	近傍	遠方	中間	全体
含有率	**97.8%**	78.9%	82.2%	86.3%

3.4.3 考察

表 4 を見ると, 実験 1 の結果 (表 1) と比較して, 近傍, 遠方, 中間の含有率が向上していることがわかる. この結果から語の階層構造と嗜好情報から算出した指標の組み合わせは, 話題選択の精度向上に対して有効であると判断する.

4 おわりに

本論文では人間と計算機との会話を目標にし, 話題選択の指標として, Wikipedia から構築した語の階層構造と Twitter から抽出した嗜好情報を用いる手法を提案した. 語の階層構造から算出した指標は候補語群を 3 つに大別することがわかり, 各グループの特徴を調査を行った結果, 近傍グループからは先行研究に比べ高い精度が得られることがわかった. また近傍グループは継続語, 遠方グループは発展語に偏った選別する性質があることがわかり, 継続語, 候補語の選択の指標として使用できる可能性を見出した. また嗜好情報から得られた指標とユーザー数との影響を調べた結果, ユーザー数は含有率に影響することがわかった. また n_{ab} がユーザー数に依存してしまうこと, ユーザー数が多い場合に ER の方が良い結果であることから, 候補語群は e_{ab} を用いた ER の方が適当であると判断した. さらに語の階層構造と嗜好情報の指標を組み合わせた結果, 話題選択の精度向上に有効であることがわかった.

今後の課題としては, 語の階層構造の構築における精度の向上, 規模を拡大した評価実験を行い詳細な評価を行うこと, 既存の会話文生成のアルゴリズムと組み合わせて実際に会話を行う実験を実施するである.

謝辞

本研究の一部は科研費（23720222）の助成を受けたものである. また IBM 社のアカデミックイニシアティブプログラムの支援を受けたものである.

参考文献

[1] 奥村紀之, 土屋誠司, 渡部広一, 河岡司. 概念間の関連度計算のための大規模概念ベースの構築. 自然言語処理 = Journal of natural language processing, Vol. 14, No. 5, pp. 41–64, 2007-10-10.

[2] 林輝大, 奥村紀之. 6x-7 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 (対話, 学生セッション, 人工知能と認知科学, 情報処理学会創立 50 周年記念). 全国大会講演論文集, Vol. 72, No. 2, pp. 2–629 – 2–630, 2010-03-08.

[3] 柴木優美, 永田昌明, 山本和英. 日本語語彙大系を用いた wikipedia からの汎用オントロジー構築. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2009, No. 4, pp. 1–8, 2009-11-09.