様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　６月１０日現在

機関番号：５６２０３
研究種目：若手研究（Ｂ）
研究期間：２０１２～２０１３
課題番号：２３７２０２２２

研究課題名（和文）
言語の非線形性に基づく自然な対話を実現するための発話モデルの構築

研究課題名（英文）
A Construction of Automatic Utterance Model based on Non-linearity of Natural Language for Realizing Natural Conversation between Humans and Computers

研究代表者
奥村　紀之（NORIYUKI OKUMURA）
香川高等専門学校・情報工学科・助教
研究者番号：４０５１０２７７

研究成果の概要（和文）：
本研究では，言語の非線形性に着目した発話モデルの構築を行った．特に，発話を行う際に重要となる話題の選定と，話題選定の基盤となる連想システムに関する研究を行った．話題の選定については，Twitter のログデータと Wikipedia から構築した語の階層構造を用いることで人間らしい話題選定を 90%の高い精度で可能とし，連想システムに関しては，これまで人手で精錬していた概念ベースをほぼ自動的に構築できる見通しを立てることに成功している．

研究成果の概要（英文）：
In this research, we constructed a automatically utterance model based on the non-linearity of lingua. Especially, we proposed the selection method of topics and the association system that provided the basis of topics selection. We enabled the topic selection with about 90% accuracy using Twitter’s log data and the hierarchy knowledge base using Wikipedia about topics selection. We estimated the automatically construction method of Concept-base.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,400,000	1,020,000	4,420,000

研究分野：人文学
科研費の分科・細目：言語学
キーワード：話題選定・連想システム

１．研究開始当初の背景

言語表現は本質的に非線形性を有しており，言語処理においては避けて通れない問題であることが明らかにされている．本研究の中核となる連想知識(概念ベース)においても，線形性を有するものであるため，人間らしい思考・思案を表現する上で障壁となっていた．そのため，非線形性を考慮した連想システムおよび発話モデルの考案が求められていた．

２．研究の目的

本研究の目的は，計算機が人間のように能動的に発話を生成するためのプロセスをモデル化することである．人間は刻一刻と変化する対話の中で適宜情報を取捨選択し，その状況に応じた話題を抽出し適切な発話を行っている．特に，単一の話題について対話を行っている場合でも，状況(話し相手)に応じて人間は多様な話題を提供する能力を有している．これまでの対話モデルの研究では会話の流れに即した応答を行うものが一般的であった．本研究では，計算機が対話相手の情報を適宜収集・整理し，対話を単一の話題だ

けでなく発展性のあるものへ展開するための手法を構築し，常に人間の側から新規の話題を提供し続けなければならないという問題を解決する．

３．研究の方法

本研究では大きく分けて２つの検証を進めてきた．一つ目は，コンピュータと人間が円滑にコミュニケーションを行うための話題選定に関する研究，もう一つは話題選定の基準となる概念ベースの構築である．それぞれの研究方法について述べる．

３．１　話題選定に関する研究

話題選定に関しては，人間が日常的に行っている会話に着目し，話題を以下の三種に分類した．

・継続的な話題

現在会話している内容をそのまま継続させる話題．

・発展的な話題

現在会話している内容と，過去に会話してきた内容，あるいは対話相手に関する情報を基に，話題を転換するような話題．

・新規の話題

対話相手に対して，まったく新しい話題を提供する場合．

これらの話題に対して，現在の対話内容の話題と，これから提示すべき話題の内容の関連の強さを定量化することによって，適切な話題を提示するための基準について検討を行った．

話題語同士の関連の強さの定量化には，Wikipedia と日本語語彙大系から構築した語の階層構造を用いている．特に，話題語を選定する上で，Wikipedia のような大規模辞書情報は非常に有用であるが，Wikipedia を階層構造として捉えた場合，最上位ノードが複数存在してしまうという問題が起きる．そこで，日本語語彙大系を Wikipedia のアッパーノードとして利用することにより，一意に決定できる最上位ノードから，単語間の距離を産出することができる．

この階層構造を用いた語と語の距離推定により，各話題語がどのような分布になっているのかを検証している．

３．２　概念ベースの構築

従来，概念ベースは半自動的に構築され，適宜人手によって精錬作業が施されてきた．しかし，人手による精錬を行うため，再現性に乏しく，理論的に説明できない要素を多数含んでいた．

そこで，話題選定のための基準として新たに概念ベースを構築し，再現性を確保するとともに性能向上を目指した．

本研究では，EDR の概念辞書を素材として，見出し語を概念とし，説明文に含まれる自立語を概念を特徴付ける関連語群(属性)として定義した初期段階の概念ベースに対し，連想語群を拡張する手法について検討した．

特に，概念ベースに定義されている概念は，属性として保持する語もまた概念として定義されているという特徴を持っており，属性を連鎖的に取得することができる．初期概念ベースの段階では，属性連鎖により何次の属性集合で飽和するかを検証した．また，飽和した状態において，サンプル概念の属性をすべて目視で評価し，獲得可能な属性，他手法が必要となる属性を検討した．

検討結果により，概念ベースを見出し語と説明文から単純に構築しただけでは取得できないような連想語群が多数存在することがわかり，他の手法を導入することが必須であることがわかった．そこで，近年着目されているテキストマイニングの技術を用いることによって，取得できなかった属性を取得可能であるかどうかの検証を行った．検証には，初期段階の概念ベースと同様，EDR の概念辞書を利用し，テキストマイニング装置として IBM 社の Content Analytics を利用し，解析を行った．

４．研究成果

本研究は大きく分けて３．１と３．２に述べた２つの成果がある．それぞれについて成果を述べる．

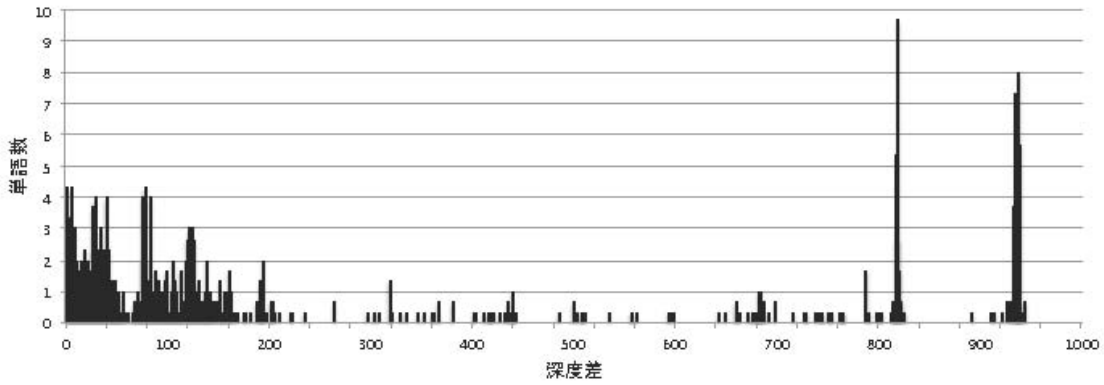
４．１　話題選定に関する研究

話題選定に関する研究では，現在話している話題として「勉強」「動物」「旅行」を想定し，次に提示すべき話題が適切であるかどうかの評価を行った．評価には，上記３種類の話題に対し，それぞれ概念ベースを用いて二次属性まで取得し，その中から適切と思われる話題抽出を行った．

話題選定基準として Wikipedia と日本語語彙大系から構築した語の階層構造を用いており，階層構造における深度差(最上位ノードからの深度の差)を用いている．図に示すように，話題候補となる語群と，現在の話題の間に深度差として 0～400 までの近傍グループ，400～800 までの中間グループ，800 以上の遠方グループに分けられることがわかった．なお，この深度差による選別については，T 検定，F 検定により優位な差が確認されている．表１に，各グループにおける適切な話題の含有率(取得された集合内に適切な話題語が含まれている率)を示す．

表 1. 各グループにおける含有率

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



また,各グループにおけるF値を算出した.表2に結果を示す.

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

表2および図より，継続した話題を抽出するためには，近傍グループ，中間グループに所属する話題語を提示することが必要であり，発展的な話題を抽出するためには，遠方グループに所属する話題語を提示する必要があることがわかった．これにより，人間とコンピュータが円滑にコミュニケーションを図る際に重要となる話題選定の基準を構築することができた.

４．２　概念ベースの構築

概念ベースの構築については，まず連鎖属性の傾向調査から行った．特に，概念ベースに定義される概念は，見出し語とその見出し語から常識的に連想される特徴語群である属性の対の集合として定義される必要がある．そこで，EDR の概念辞書から構築した初期の概念ベースを属性連鎖展開し，何次属性まで展開した際に収束するのか，収束した段階での属性の傾向はどのようになっているのかを検証した．結果を図2に示す.

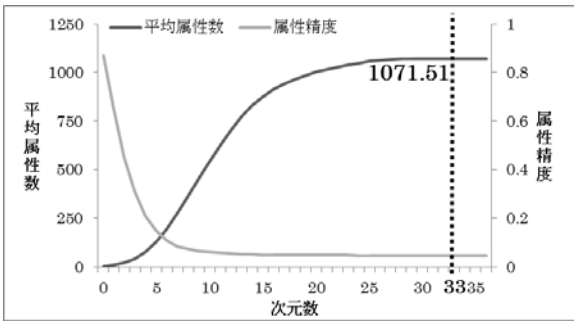


図2. 連鎖属性と精度の変化

図2に示した連鎖属性と制度の変化は，サンプル概念として100を準備し，それぞれの属性の傾向調査を目視で行ったものである.

各概念について調査したところ，33次元までの属性連鎖によって概念ベースは収束することがわかった．このときの属性の精度はきわめて低く，ほとんどが概念と関係のない属性が取得されている.

一方で，適切と思われる属性も平均して30～40語は含まれていたが，連鎖させるだけでは取得できないような連想語も多数考えられる．たとえば，「トマト」に対する連鎖属性の中で適切と思われる語は，「果実，食べる，実，料理，植物」などが取得される．しかし，「トマト」から常識的に連想される「ソース，赤，ペースト」といった語は，連鎖させるだけでは取得不可であることが判明した.

そこで，テキストマイニング装置を利用することにより，EDR の概念辞書を相関分析することによってどのような集合が得られるかを検討したところ，上記の「ソース，赤，ペースト」に加え，「レタス，はさむ，ポテト」といった，トマトから常識的に連想可能な語群が，同じ辞書の素材を別の観点から分析することにより取得可能であることがわかった.

このことから，今後は概念ベースの自動構築および精錬を実施し，より人間らしい対話コンピュータを実現することを目指す.

４．３　その他の特筆すべき成果

本研究の応用実験として行っていた顔文字に含まれる感情成分に基づく感情判断システムの拡張については，学会発表 4. にも示しているように，情報処理学会第 75 回全国大会での研究成果報告を行い，学生奨励賞を受賞している．当該研究では，概念ベースの応用事例として研究を進めている常識的判断システムのひとつである感情判断システムを顔文字と組み合わせることで拡張し，より人間らしい感情推定を行うシステムを提案している.

特に，包括的に顔文字について調査した研究はまだ少なく，文章から連想される感情と顔文字から連想される感情を組み合わせることで，発話者の心情を推定する，また顔文

字そのものを形態素とみなした解析を行うなど，非言語情報を適切に扱う上で重要となる研究を開始できている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１７件）

1. Hiroki Takimoto, Noriyuki Okumura and Masashi Mizuno：「Time Line Inference from Story Sentence with Time Series」, SCIS-ISIS 2012, pp.1089-1093, 2012 年 11 月
2. Noriyuki Okumura and Yuto Hatakoshi：「A Refining Method of Obtained Attributes to Characterize Undefined Concepts using Search Engine」, Proceedings of KEOD2011 - International Conference of Knowledge Engineering and Ontology Development, pp.493-497, 2011 年 10 月
3. 奥村紀之, 豊嶋章宏:「語の相関関係を考慮した概念ベースの連想語取得手法の検討」, 言語処理学会第 19 回年次大会, P4-2, 2013 年 3 月
4. 大西智佳, 奥村紀之:「顔文字に含まれる感情成分に基づく感情判断システムの拡張」, 情報処理学会第 75 回全国大会, 6R-4, 2013 年 3 月
5. 楠和馬, 奥村紀之:「話者の負担を考慮した話者識別と音響モデルの検討」, 情報処理学会第 75 回全国大会, 5T-3, 2013 年 3 月
6. 豊嶋章宏, 奥村紀之:「概念ベースにおける属性連鎖の傾向と属性集合の評価」, 情報処理学会第 75 回全国大会, 1Q-3, 2013 年 3 月
7. 松岡雅也, 奥村紀之:「統計的手法を用いた命題的知識の真偽判断システム」, 情報処理学会第 75 回全国大会, 4Q-1, 2013 年 3 月
8. 奥村紀之, 大西智佳:「文字情報と顔文字からの話者感情推定」, 信学技報、Vol.112, No.268, NLC2012-30, P31-33, 2012 年 10 月
9. 奥村紀之:「Web を利用した概念ベースの自動構築」, 言語処理学会第 18 回年次大会, P3-3, 2012 年 3 月
10. 林輝大, 奥村紀之:「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」, 言語処理学会第 18 回年次大会, D2-4, 2012 年 3 月
11. 西村太佑, 奥村紀之:「マルコフテーブルによる学習を用いた辞書型対話システム」, 情報処理学会第 74 回全国大会, 5Q-1, 2012 年 3 月
12. 野崎徹郎, 奥村紀之:「単語の単位度比較を用いた文章簡略化システム」, 情報処理学会第 74 回全国大会, 1V-1, 2012 年 3 月
13. 町田啓悟, 奥村紀之:「感情判断に基づく物語文章からの感想文自動生成手法」, 情報処理学会第 74 回全国大会, 3R-1, 2012 年 3 月
14. 奥村紀之:「Web の時間的変化に着目した未定義概念の属性精錬手法」, 言語理解とコミュニケーション研究会（NLC), 信学技報, Vol.111, No.228, NLC2011-26, 23-26, 2011 年 10 月
15. 奥村紀之:「言語の非線形性に着目した連想システムの構築」, NLP 若手の会第 6 回シンポジウム, 2011 年 9 月
16. 林輝大, 奥村紀之:「発話のための Web を用いた背景的知識の構築手法」, FIT2011, 2011 年 9 月
17. 瀧本洋喜, 奥村紀之:「物語文章における時系列情報の抽出」, 信学技報, Vol.111, No.119, NLC2011-1, pp.1-6, 2011 年 7 月

〔その他〕
ホームページ等
http://www.di.kagawa-nct.ac.jp/~okumura/index.html

６．研究組織
(1)研究代表者
奥村紀之（OKUMURA NORIYUKI）
香川高等専門学校・情報工学科・助教
研究者番号：４０５１０２７７