1C5-01

# 対話文を用いた語の関連性の構築手法の提案

How to construct word relationship using conversation text corpus

古田 陽 Akira Furuta　　榑松 理樹 Masaki Kurematu　　藤田 ハミド Hamido Fujita

岩手県立大学
Iwate Prefectural University

We propose a method of constructing word relationship using conversation text corpus. To begin with, this method extracts words from text corpus. Next, it calucurates relationship between extracted words based on the distance between two words and the number of times which appeared. Finaly, it constracts concept hierarchy using word relationships. In order to evaluate the effciently of this method, we compared word relationship made by a prototype system with handmade word relationship.

## 1. はじめに

人とコンピュータとのより自然なインタラクションの実現をめざし，対話システムの研究が行われている．対話システムを実現するには，多くの知識が必要であり，その一つとして対話の進め方に関する知識があげられる [tanaka]．対話の進め方に関する知識は，人の対話から得ることができる．人の対話に注目すれば，人は，先の発言中の語と関連する語を以後の発言において利用することが多いことがみてとれる．我々は，この点に注目し，これを「語の関連性」として捉える．すなわち，本稿で考える語の関連性とは，語の類義性に基づくものではなく，対話文中において，ある語を受けて別の語が利用される度合い（利用傾向）を示すもの，対話の進め方に関するより具体的な知識である．この関連性を用いることにより，対話における語の利用におけるつながりがより自然なものとなり，対話システムは，人との対話をよりスムーズに行うことが期待できる．

しかし，語の関連性を人手で構築することは，語彙量から考えても困難である．この問題点に対し，本稿では近年研究資料として収集され利用可能となってきている対話文テキストコーパスを用いた語の関連性構築手法について提案する．本手法では，コーパスから半自動的に語の関連性を構築を行う．このことから，構築時間や人の作業料の軽減が期待できる．さらに我々は，本手法に基づいた実験システムを構築し，手法の有用性の評価を行った．

## 2. 語の関連性構築手法

語の関連性構築手法の概要を図 1 に示す．本稿で述べる語の関連性とは，対話において、ある語を受けて別の語が利用される度合いを示すものである．この度合いを関連度と呼び，語の関連性構築においては，任意の二つの語間の関連度を求める．ある語 A を受けて別の語 B が利用されるとは，対話文中においては，語 B が語 A の後に現れるという形で確認することができる．その回数が多ければ多いほど，その利用方法は一般的なものであると考えることができる．しかし，ある語の後に現れたとしても，その間が離れていれば，語 B が語 A を受けて利用されている可能性は低い．以上の事から，本手法では，語と語の出現距離（語間の語数）と一定の距離内での出現回数に注目し，語の関連度を求める．また複数の語に対する上位概念を導入することで，より多くの場合への適用が可能となることから，語の関連性に基づく上位概念の抽出を試みる．以下，提案する構築手法について，入出力，処理手順と分けて説明を行なう．

連絡先: 連絡先:古田陽 (a-furuta@jaist.ac.jp)，現所属：北陸先端科学技術大学院大学

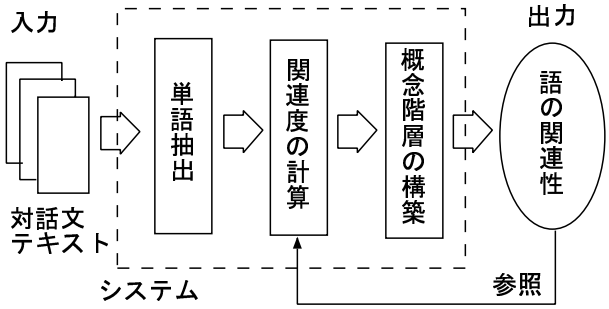


図 1: 語の関連性構築手法

### 2.1 入出力

入力として，対話文テキストコーパスを与える．またここで与える各対話文テキストは，ある一つの話題に対する人と人との 1 対 1 の対話文テキストであり，各発言の発言者が明確にわかっているものとする．

出力となる語の関連性は，語の集合および，語の集合に含まれる任意の二つの語間の関連度または上位下位関係から構成される．

### 2.2 処理手順

構築手法は，(1) 単語抽出，(2) 関連度の計算，(3) 概念階層の構築，の順で行う．以下，それぞれの処理工程について説明する．

(1) 単語抽出

単語抽出工程においては，各対話文テキストから語の関連性構築の対象となる語および関連度算出に必要な値を求める．

(1.1) 対象語抽出

はじめに対話文テキストを形態素解析し，形態素列へと変換する．次に，形態素解析列から，名詞，固有名詞，および用言を語の関連性の構築対象として抽出する．これら抽出した語を，本稿では，以後，対象語と称す．また対象語抽出においては，複合語は，それらを構成する語単位でのみ取り出し，複合語は対象としない．

(1.2) 近傍度の計算
抽出した対象語に対し，任意の対象語の組に対する近傍度を計算する．

近傍度とは，発話文中における語間の出現箇所の近さを示した値である．この値が大きいほど，一方の語がもう一方の語を受けて利用されると考える．

対話文テキスト $D$ 中における単語 $wa$ と単語 $wb$ の近傍度 $ne$ は，式 1 にて求める．式 1 において，$N$ は $D$ における $wa$ の出現総数，$wa_i$ は $wa$ の $i$ 番目の出現位置（先頭から数えて何個目の $wa$ が，全体の何番目の単語か），$M$ は，$wa_i$ を基準とした有効範囲内に出現する $wb$ の個数，$wb_{i,j}$ は，$wa_i$ の有効範囲内に出現する $wb$ をそれぞれ示す．また，$wa_i$ と $wb_{i,j}$ の距離とは，その間に出現する対象語の個数を意味し，有効範囲は，ユーザが指定する．

$$ne(wa, wb, D) = \sum_{i=1}^{N} \frac{1}{\sum_{j=1}^{M}(wa_i \text{と } wb_{i,j} \text{の距離} + 1)} \quad (1)$$

(1.3) 出現数の計算
次に任意の対象語の組に対する出現数を求める．

出現数とは，対話文中において対象語の組が一定距離内に共に出現した回数を示す．本稿では，この値を利用された回数と見なし，この値が多いほど利用されやすいと考える．出現数は，近傍度が計算されるたびに一定値加算する．

(2) 関連度の計算

(2.1) 近傍度，出現数の計算
各対話文テキストに対して求めた対象語間の各組の近傍度，出現数の合計をそれぞれ求め，それを入力された対話文テキストコーパスにおける対象語間の近傍度，出現数とする．

(2.2) 既存の語の関連性へのフィードバック
対話文テキストコーパスにおける対象語間の近傍度と出現数を，それぞれ同一対象語間の既知の近傍度と出現数に加算する．

(2.3) 出現修正値の適用
(2.2) までの工程で得た近傍度は，対話文テキストコーパスによる偏りが大きいことが予想される．この問題に対し，本稿では，近傍度に対して出現修正値を与えることで解決をはかる．出現修正値は 0 から 1 の定義域内の実数であり，フィードバック処理後に，入力対話文テキストコーパスにおいて，一定の出現数を下回った対象語間の近傍度に出現修正値を掛けたものを新しい近傍度とする．

(2.4) 関連度の計算
以上の工程で得られた近傍度および出現数から，式 2 によって関連度を求める．式 2 において，$wa, wb$ は対象語を，$h(wa, wb)$ は，$wa, wb$ の組の出現数をそれぞれ示す．また、本式において $wa$ は辞書順において $wb$ の前の語であり，$wa$ と $wb$ の関連度は，$wa$ の後に $wb$ が出現した場合の近傍度，出現数とその逆の場合の近傍度，出現数のそれぞれの合計値を利用して求める．

$$\begin{aligned} &\text{関連度 } r(wa, wb) = \sqrt{(Ne)^2 + (\log_{10}(H)^2} \\ &Ne = ne(wa, wb) + ne(wb, wa) \\ &H = h(wa, wb) + h(wb, wa) \end{aligned} \quad (2)$$

関連度は，対象語が対話文中で近くに出現することが多いほど，高い値となる．式 2 において，出現数の常用対数をとる理由は，関連度に対する近傍度に対する影響，特に出現数が多くなった場合での影響を押えるためである．

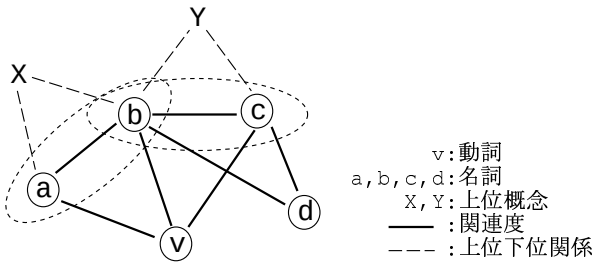


図 2: 概念ペアの抽出

(3) 概念階層の構築
最後に得られた対象語間の関連度に着目し，概念階層を構築する．概念階層は，(3.1) 概念ペアの抽出，(3.2) 任意の概念ペアに対する上位概念の決定，(3.3) さらに上の上位概念の構築，(3.4) 上位概念の関連度の計算の順で構築する．

(3.1) 概念ペアの抽出
はじめに，対象語に含まれる動詞 v と，v と関連度の高い名詞および固有名詞をすべて抽出する．抽出した名詞，固有名詞からなる 1 対 1 のすべての組合わせのうち，関連度が存在する組を概念ペアとする．以後，抽出した概念ペアの上位概念の抽出を行う．概念ペアの抽出方法の模式図を図 2 に示す．図 2 において，名詞 a,b,c は，動詞 v と高い関連度を持つ．これらの総あたりのペアのうち，ペアを構成する語間に関連度が無い (a,c) のペアを除いた (a,b)，(b,c) を概念ペアとして抽出する．抽出した概念ペアに対し，上位概念の抽出を試みる．

抽出した概念ペアの上位概念の抽出を行う理由は，次の通りである．動詞 v に対して関連度の高い名詞，固有名詞は，動詞 v を受けて利用することが多い名詞，固有名詞とみなせる．すなわち，それらは同一の動詞に対して同じように利用されるものとなり，同一のカテゴリ（上位概念）に属する可能性が考えられる．よって，それらの上位概念の抽出を試みる．

(3.2) 任意の概念ペアに対する上位概念の決定
抽出した概念ペアを構成する名詞または固有名詞のペアをユーザに提示し，その上位概念の有無をユーザが決定する．上位概念が存在すると判断した場合，ユーザが，上位概念の名前を与え，それを概念ペアを構成する語の（共通の）上位概念とする．このとき上位概念名が語の関連性に含まれていれば，その語の下位概念として，概念ペアを構成する語を登録する．

(3.3) さらに上の上位概念の構築
(3.2) で獲得した上位概念の上位概念が存在する可能性があるため，本稿では，次の 2 つの場合に限り，さらに上位概念の抽出を試みる．

1 つめは，2 つの上位概念が下位概念の一部を共有する場合である．下位概念の一部を共有するということは，上位概念の間に共通する部分があると判断し，その共通部分に対応する概念が存在すると考える．その概念に対し，ユーザに質問を行う．

2 つめは，上位概念 X の下位概念 A と上位概念 Y の下位概念 B の間に関連度が存在する場合である．ここで，A は Y の下位概念ではなく，また，B は X の下位概念ではない．この場合の A と B は，対話文テキストによって上位概念は見つからなかったが (3.2) で捉えた場合と同じと考え，概念抽出を試みる．

以上の操作を上位概念が作り出せなくなるまで繰り返し行う．

(3.4) 上位概念の関連度の計算
上位概念に対しては，その下位概念に含まれず，下位概念すべ

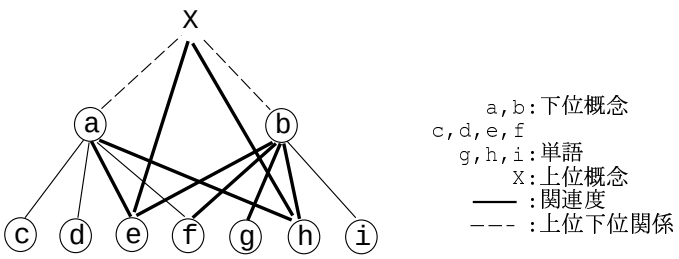


図 3: 上位概念の関連度の計算

てと関連度の高い語との間に関連度を与える．対象となる語が下位概念すべてと関連度が高いということは，それら下位概念に共通の特徴（関連性がある）であり，それは上位概念が持つべき特徴であるという考えに基づいている．図 3 に，この考えの模式図を示す．図中において，関連度の線の太さは関連度の高さを表している．図 3 において，a と b の上位概念である X は，a，b どちらともと高い関連度をもつ e と h のみに対し，関連度を与える．

上位概念 $X$ と概念 $t$ との関連度については次の方法で計算する．ここで概念 $t$ は，$X$ の下位概念すべてと高い関連度を持つ語である．既存の語の関連性において与えられている $X$ と $t$ の間の関連度を $r_0$ とする．関連度が存在しない場合は，$r_0$ は 0 とする．この値に，$X$ の下位概念全てと $t$ との関連度の平均を加えたものを，$X$ と $t$ の間の関連度とする．計算式を式 3 に示す．なお，式 3 において，$C_i$ は，$X$ の下位概念，$L$ は $t$ と関連度を持つ $X$ の下位概念数，$r(A,B)$ は，$A$ と $B$ の関連度 ($A$ は $B$ より辞書順で前にくる) を意味する．

$$\text{関連度 } r(X,t) = \frac{\sum_{i=1}^{L} r(C_i,t)}{X \text{ の下位概念の総数}} + r_0 \qquad (3)$$

## 3. 評価実験

本提案手法の妥当性を評価するために，本手法に基づくシステムのプロトタイプを実装し，その評価実験を行った．なおプロタイプシステムの単語抽出においては，icot 形態素辞書 [www.icot.or.jp] を利用している．

### 3.1 実験概要

本実験では，システムが構築した語の関連性の妥当性を評価することにより，手法の有用性の評価を行う．語の関連性の妥当性の評価基準には，システムで構築した語の関連性の一部と，それらを構成する語に対し，被験者が作成した語の関連性との相関係数を用いる．相関係数を用いる理由は次の通りである．本システムにおいて構築する語の関連性は，語の利用傾向（想起しやすさ）を示すものである．よって，システムで得た関連性と人が判断した関連性との傾向が類似していれば，システムが得た語の関連性は人手によるものと類似していると考えられ，対話時に有用であると考えられる．そして手法の出力が有用であれば，手法も有効であると判断する．

### 3.2 実験結果

(1) システムによる語の関連度の構築

システムに対する入力データとしては，北九州立大学国際環境工学部で公開されている対話文サンプルコーパス [www.env.kitakyu-u.ac.jp]，および，個人で収集した対話文サンプルコーパスの計 101 個を使用した．これらのコーパスは，1 ファイルごとに 1 つの話題について対話されており，その話題は，日常的なものである．

本実験において，近傍度の有効範囲としては，$wa_i$ から $wa_{i+1}$ の間または，$wa_i$ からの距離（語数）が，10 個以内とする．ただし，用言については，文を越えないものとする．また，出現数の計算における加算値は，1 とする．出現修正値は，対話文テキストコーパスにおいて，出現数が 0 だった場合に，0.8 を与える（0.8 倍する）．

本システムでは，コーパスから，4759 個の語を抽出し，それらの組合わせのうち，323063 組の関連性（関連度）を得た．さらに，231 個の上位概念を構築し，それら上位概念を含めた関連度を持つ概念ペアは，323088 組であった．なお今回構築された上位概念のうち 36%は語の関連性に存在するものであった．

(2) 人手による語の関連度の構築

次にシステムが構築した語の関連度との比較対象となる，人手による関連度を次の方法で構築した．

システムで抽出された語のうち，「学生，アルバイト，勉強，仕事，映画，アメリカ，時給，行く，聞く，見る」の 10 語を構築対象とする．これらの語は，システムへの入力である対話文テキスト中から，話題のキーとなっている語を実験者が選出した．

次に 28 名の被験者が，それらの語の関連性（利用傾向）について，4 段階評価を行った．なお利用する傾向が強いものに対して高い値を，弱いものについては低い値を与える．この時，語の出現順（A の後に B が利用するか，B の後に A を利用するか）も意味があると考え，別々に値を与えた．以上によって与えられた語の各組み合わせに対する関連度の平均から構成されるものを，人手による語の関連度とする．なお今回の実験における被験者の内訳は，10 代男性:2 名，10 代女性:1 名，20 代男性:22 名，20 代女性:2 名，30 代男性:1 名である．

(3) 相関係数の算出

得られた二つの語の関連性に対して相関係数を求めた結果を表 1 に示す．表において，MAX とは全組み合わせに対する相関係数が最も良かった場合，MIN とはその逆，AVE とは両方の平均をとった場合を意味する．

表 1: 実験結果

| | 組み合わせ | | | |
|---|---|---|---|---|
| 観点 | 名詞, 名詞 | 名詞, 動詞 | 動詞, 動詞 | 全て |
| MAX | 0.63 | 0.788 | 0.862 | 0.657 |
| MIN | 0.484 | -0.015 | 0.698 | 0.271 |
| AVE | 0.557 | 0.441 | 0.791 | 0.489 |

### 3.3 評価・考察

表 1 で示す通り，得られた相関係数にはばらつきがある．(動詞，動詞) は，全体として高い値を得ている．このような結果を得たのは，今回対象とした動詞の数が少ないこと，ある動詞を受ける動詞は比較的限定されていることが大きな理由と考えられる．(名詞，名詞) は，やや高い値を得た．このような結果を得たのは，今回対象とした名詞の利用範囲（対話文の話題）が，近いことが考えられる．(名詞，動詞) は，ばらつきが生じた．このような結果を得た理由としては，語の利用の方向性が考えられる．すなわち，ある名詞を受けて利用する動詞は比較的限定できるが，動詞を受けて利用する名詞は，対話の話題などその利用する状況に依存するところが大きく，被験者がどのようにその状況を考えたかで大きく変化する．アルバイトや

時給というものは，その利用範囲が限定されやすいが，学生，アメリカなどは利用範囲が広く，特に主語になりえる言葉において，そのばらつきが大きく現れている．実際，被験者の結果においては，(名詞, 名詞）(動詞, 動詞）はばらつきが少なく，(名詞, 動詞）はばらつきが大きい．

以上の実験結果から，本手法は，同一品詞間についてはある程度有用な関連性を抽出することができるが，品詞が違う場合については，不十分であると評価できる．このような結果を得た大きな理由としては，利用順を反映していないこと，対話環境による変化を考慮していないことが考えられる．よって関連性構築手法に対し，利用順や対話環境を考慮する仕組みを加えることによって，より精度の高い関連性を得ることが考えられる．

また今回の実験において，システムが抽出対象とした対話文は，一つの話題を対象としているが，人手によるものは，話題を限定せず，対話中ではない．この点を考慮し，今回の提案手法においては，テキストコーパスによる偏りを防ぐ目的で出現修正値を導入したが，十分な効果を発揮できなかった．語によっては，利用される状況によって大きな変化が生じることから，それらの語に対する対応，すなわちテキストコーパスによる偏り，作成状況の違いを埋めるなんらかの改善策が必要である．その一つとして，観測されない性質に対して考慮する最大エントロピー法の利用が考えられる．

一方概念構築においては，今回提示された概念ペアの大半に対しては，ユーザが上位概念を思いつかなかった．この点から，本手法の概念抽出の精度は低い．このような結果を生じた大きな要因としては，上位概念を構築する観点とそれを見つけるための観点の違いが大きく影響をしていると考えられる．その点から，利用時における語順や動詞に対する意味の役割の利用など新たな観点を追加するとともに，電子化辞書などの既存の情報リソースを利用することが考えられる．

今回はシンプル方法による語の関連性構築手法を提案したが，語の性質や対話環境による変化が要因となり，十分な成果を得られていない部分が見受けられた．今後は，現在の手法に対し，このような点への改善方法を検討する必要がある．

## 4. おわりに

本稿では，人との対話を行う対話システムの実現に必要な知識として，語の関連性（利用傾向）に注目し，それを対話文テキストコーパスから半自動的に構築する手法について提案した．提案手法は，対話文の中において，近くに出てくる回数の多い語間には高い関連性があると考え，コーパスにおける語の出現位置，回数からから，語の関連性の構築を試みる．さらに得られた語の関連性に基づき，概念階層の構築を試みる．

本手法の有用性を評価するために，プロトタイプを構築し，評価実験を行った．実験では，システムが構築した語の関連性と，人手により構築された語の関連性との比較することによりシステムの有効性を評価した．評価基準として相関係数をもとめった結果，(動詞と動詞）で最大値を得，(名詞と動詞）では最小値を得た．全体としては，0.489 であった．この結果，本手法は，利用順や対話環境の影響が少ない語の関連性に対しては有効であるが，利用順や対話環境の影響が大きい語の関連性に対しては，有効ではなかった．今後の課題としては，現在考慮していない利用順や対話文の解析結果に基づく対話環境を手法に反映させることで，手法の能力の向上があげられる．

## 参考文献

[tanaka] 田中穂積: 自然言語処理 - 基礎と応用 - , 電子情報通信学会,(1999)

[www.env.kitakyu-u.ac.jp] Hypermedia Corpus of Spoken Japanese (『平成８－１０年度文部省科学研究費補助特定領域研究「人文科学とコンピュータ」公募研究（「日本語会話データベースの構築と談話分析」研究代表者 上村隆一）の成果による』) :http://www.env.kitakyu-u.ac.jp/corpus/

[www.icot.or.jp] icot フリーウェア :http://www.icot.or.jp/ARCHIVE/HomePage-J.html