2G3-01

概念的関連性に基づく雑談の話題転換点分析

An analysis of topic changes in free conversation using conceptual relations.

藤本 英輝[*1] Eiki Fujimoto　高梨 克也[*2] Katsuya Takanashi　河野 恭之[*1] Yasuyuki Kono　木戸出 正継[*1] Masatsugu Kidode

[*1]奈良先端科学技術大学院大学 Nara Institute of Science and Technology

[*2]独立行政法人 情報通信研究機構 National Institute of Information and Communications Technology

This paper investigates the feature of topic changes in three-person free conversation by analyzing conceptual relations between an utterance and the topic in which the utterance is performed. Conversations are often inactivated when participants have no knowledge about each other and cannot find any suitable topic. In such a scene, if a system can introduce some suitable topics to participants in the conversation, it will activate the conversation. We analyzed the relationship between manually classified topic changes and the utterance in free conversation corpus to find a method for automatically classifying topic changes.

1. はじめに

コンピュータネットワークの普及・拡大に伴い，人々の出会いの機会は格段に増えた．しかし初対面の人間同士の対話では次に提出する話題が見つけられず対話が停滞することがある．このような場面において，対話参加者に対しシステムから適切な話題を提供できれば，再び対話を活性化させることが可能だと考えられる．一般的な対話では話者は聴者にある程度の予備知識があることを前提に，それを暗黙の了解として踏まえ新しい事柄を伝える [1]．雑談の場合も新規話題の提出タイミングやその内容についての自由度は高いが，それまでの会話の流れと無関係な話題に遷移することは少なく，直前の話題と関連性をもった話題が選ばれる．このことから，対話支援システムには話題の流れを考慮し自然な繋がりをもつ次話題を選択する能力が必要であると考えられる．話題の繋がりの自然さを測る指標として，話題・発話間の概念的な関連性の利用が考えられる．

以上のことから本稿では，人間同士の三者雑談における話題・発話間の繋がりの自然さに着目し，概念的関連性に基づく話題転換の特徴分析を行った結果について述べる．

2. 対話コーパスの主観的分析

話題転換点における話題と発話の概念的関連性の特徴を分析するには，人間同士での雑談中に起こっている話題転換について知る必要がある．そこで対話内容に制限のない三者対話のコーパスを用い，話題転換点数の調査および話題転換の自然さによる分類を行った．

2.1 三者対話コーパスの概要

本研究では三者会話コーパス [2] を用いて分析を行った．被験者は全員大学生で，同性三人を一組としている．本稿では組み合わせが (a)AB は友人，AC も友人だが BC は初対面 (b) 全員初対面 (c)AB は友人，C は A，B ともに初対面の三つの対話を利用する．200ms 以上の休止で区切られた単位を一発話とし，発話中に一つ以上の自立語を含むものを有効発話とする．各対話の発話数および有効発話数を表 1 に示す．

表 1: 三者対話の被験者組み合わせと総発話数

	対話 1	対話 2	対話 3
組み合わせ	AB 友人	全員初対面	AB，AC 友人
発話数	780	841	917
有効発話数	413	557	562

2.2 話題転換の自然さによる分類

新たな話題が開始された発話のことを話題転換点と呼ぶ．実際の対話では，発話の最初から新しい話題が始まる場合だけでなく，発話開始時は前の話題を受け継ぎ途中から別の話題に遷移しているように見えることもある．このような場合もその発話を分割することはせず，発話全体を話題転換点とする．また同一の話題が継続されている発話区間を話題ブロックと呼ぶ．

雑談が盛り上がっているとき，実際には話題転換が起きていてもそれを明確に意識することは少ない．そこで，人間同士の雑談では現在の話題と関連性はあるが当たり前すぎない話題への遷移がよく行われ，そのような話題遷移が対話の活性化をさせやすいという仮定をおく．話題転換点発話が行われた時点での話題・発話に含まれる単語同士の概念的な繋がりの強さを指標として，話題転換の自然さを三段階に分類する（図 1）．

TypeA きわめて関連性の強い話題への転換

TypeB 中程度の関連性をもった話題への転換

TypeC 元の話題とは関連性がほとんどない話題への転換

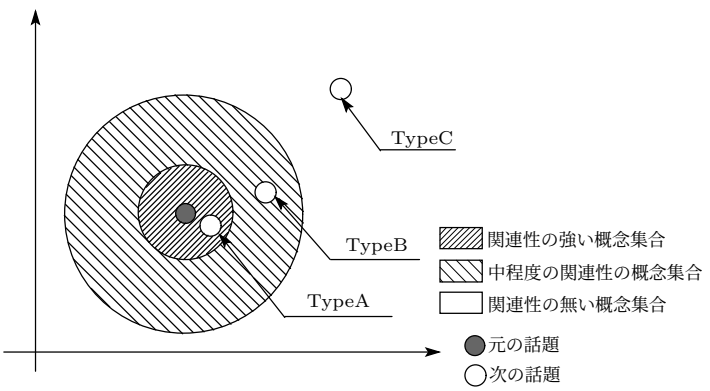

図 1: 話題転換タイプの概念図

連絡先: 奈良先端科学技術大学院大学 情報科学研究科，〒 630-0192 奈良県生駒市高山町 8916-5，eiki-f@is.naist.jp

2.3 分析結果

三つの対話から有効発話のみを対象として主観に基づき話題転換点を抽出，各話題転換点を TypeA，TypeB，TypeC に分類した．非転換点発話は TypeN とした．また，一つの話題ブロックが一定ターン以上継続した場合を対話が活性化された状態であるとし，その話題ブロックの起点となった話題転換点を活性話題転換点であるとした．あるタイプの転換が起こった際に，その話題が活性化した割合を活性化率（式 1）として分析を行った（表 2，3）．

$$\text{活性化率} = \frac{TypeX \text{ 話題転換点中の活性話題転換点数}}{TypeX \text{ 話題転換点の総数}} \quad (1)$$

表 2: 発話の分類

対話		TypeA	TypeB	TypeC	TypeN
1	総数	4	43	14	352
	活性話題	2	25	5	-
2	総数	7	58	17	475
	活性話題	3	42	10	-
3	総数	0	55	16	491
	活性話題	0	42	9	-
合計		11	156	47	1318

表 3: 活性化率

	TypeA	TypeB	TypeC
対話 1	0.5	0.58	0.36
対話 2	0.43	0.72	0.59
対話 3	-	0.76	0.56
平均	0.47	0.69	0.50

表 2 から人間同士の雑談では TypeB の話題転換が最も多用されているのが分かる．また表 3 から TypeB 発話は活性化率も高いことが見て取れる．これは 2.2 で述べた「人間同士の雑談では現在の話題と関連性はあるが当たり前すぎない話題への遷移がよく行われ，そのような話題遷移が対話の活性化をさせやすい」という仮定を肯定するものである．このことから，雑談を行うシステムが自ら新しい話題を提示する際には，TypeB の話題転換となるような話題を選ぶことで，人間らしい自然な対話継続が可能になると考えられる．

3. 拡張概念の定義

話題・発話間の概念的関連性を測るため，単語の持つ概念的広がりを表現した拡張概念を定義し，計算機で利用するための拡張概念辞書の作成手法について述べる．ドメインに依存しない一般的な概念間の関係が定義されている EDR 電子化辞書を用いた．

ある概念 w に対し，EDR 電子化辞書中から概念的関連性が得られる語を要素とする重みつき集合を定義し，それを用いて概念的関連性をみることにする．ここで，ある概念 w を基準概念，基準概念と関連性のある概念の集合を拡張概念と呼ぶ．拡張概念の生成に用いる関連性（リンクタイプ）は

- 一次属性 (Ex)：w の語義文中に出現する自立語概念
- 二次属性 (Ex2)：一次属性の語の一次属性
- 逆属性 (Rex)：w を一次属性に持つ概念語
- 兄弟概念 (Bn)：w と同一の直上概念をもつ概念
- 子概念 (Cn)：w の直下概念

の五種類である．これら関連性をもつ概念を要素とし，出現頻度に基づく重みを与えたものを拡張概念 Ec と定義する．

$$Ec = \{(c_1, w_1), (c_2, w_2), \ldots, (c_n, w_n)\}$$

ここで，c_i は概念，w_i はその重みである．

4. 発話・話題間の概念的関連性計算システム

本稿では発話や話題の概念を拡張概念を用いたベクトルで表現し，その類似度で関連性の度合いを見る．拡張概念を用いて発話・話題間の類似度を計算するシステムの概要を図 2 に示す．

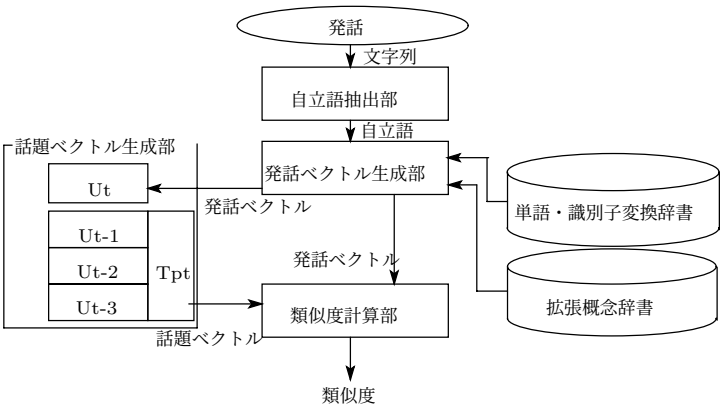

図 2: 発話・話題間の概念的関連性計算システム

文字列で入力された発話は，自立語抽出部で形態素解析され自立語のみが抽出される．抽出された自立語は発話ベクトル生成部で変換辞書を用いて単語表現から概念を表す概念識別子表現に変換され，拡張概念辞書を用いて発話ベクトル形式に変換される．話題ベクトル生成部では過去三発話分の発話ベクトルから話題ベクトルを生成し，類似度計算部において発話ベクトルと話題ベクトルの類似度が出力される．

発話中の全自立語から得られる拡張概念の和を発話ベクトルとする．いま発話から三つの拡張概念 Ec_1，Ec_2，Ec_3 が得られたとする．

$$Ec_1 = \{(c_1, w_1), (c_2, w_2), (c_3, w_3), (c_4, w_4)\} \quad (2)$$
$$Ec_2 = \{(c_2, w_5), (c_3, w_6), (c_6, w_7)\} \quad (3)$$
$$Ec_3 = \{(c_1, w_8), (c_3, w_9), (c_5, w_{10}), (c_6, w_{11})\} \quad (4)$$

これら三つの拡張概念から，発話ベクトル U は以下のように生成される．

$$\boldsymbol{U} = \{(c_1, W_1), (c_2, W_2), (c_3, W_3), (c_4, W_4), (c_5, W_5), (c_6, W_6)\} \quad (5)$$

ただし

$$W_1 = w_1 + w_8 \quad , \quad W_2 = w_2 + w_5$$
$$W_3 = w_3 + w_6 + w_9 \quad , \quad W_4 = w_4$$
$$W_5 = w_{10} \quad , \quad W_6 = w_7 + w_{11}$$

である．

過去三発話分の発話ベクトルの和を，その時点での話題の概念を表現する話題ベクトルとする．ある時点 t での話題ベクトル Tp_t は発話ベクトル U_i を用いて以下のように表せる．

$$\boldsymbol{Tp_t} = \boldsymbol{U_{t-1}} + \boldsymbol{U_{t-2}} + \boldsymbol{U_{t-3}} \tag{6}$$

ある時点の発話ベクトルとその発話が行われたときの話題ベクトルとの類似度は以下の式で定義する．

$$Sim = \frac{\boldsymbol{Tp_t} \cdot \boldsymbol{U_t}}{|\boldsymbol{Tp_t}||\boldsymbol{U_t}|} \tag{7}$$

ここで，$\boldsymbol{Tp_t}$ はある発話時点 t での話題ベクトル，$\boldsymbol{U_t}$ は時点 t での発話ベクトルである．

5. 概念的関連性に基づく話題転換の分析

5.1 話題・発話ベクトル間の類似度特徴分析

話題ベクトル，発話ベクトルはそれぞれある時点での現在の話題と発話のもつ概念を表現したものである．よって話題・発話ベクトルを構成する拡張概念が適切であるなら，同一の話題について話しているときや概念的に近い話題に遷移したときはベクトル間の類似度が高く，逆に話題が関係のないものに遷移したときは類似度が低くなると予想される．そこで 2.3 で人手認定した各話題転換点および非転換点における話題・発話ベクトル間の類似度特徴の分析を行った．

分析手法

全ての有効発話時点での話題ベクトルと発話ベクトルの類似度を計算した．それを三つの対話全てで行い，各話題転換タイプ（TypeA，TypeB，TypeC）およびそれ以外（TypeN）に分類・集計し，平均値を求めた．

結果

以下に結果のグラフを示す（図 3）．ある発話が行われた時点での話題・発話ベクトル間の類似度の平均値は，TypeA 転換点発話と非転換点発話（TypeN）が高く，TypeB，TypeC と減少しているのが分かる．これは 2.2 で示した概念的関連性の強さによる話題転換点の認定結果と一致している．しかし発話タイプ別の話題・発話間類似度の相対度数分布（図 4）を見ると分かるように，サンプル数の少なかった TypeA を除く 3 タイプが全て類似度 0.1 の階級の相対度数が最も高く，類似度が高くなるにつれて相対度数が減少している．3 タイプの平均値の差は，類似度が 0.5 を超える発話の相対度数の差に起因していると考えられる．このため平均値にはタイプごとの特徴が現れるが，話題・発話ベクトル間の類似度のみを用いてある発話が行われた時点で発話タイプを特定することは困難である．

この原因は二つ考えられる．一つは単語を概念識別子に対応付ける際の曖昧性である．一般に一つの単語に対して，同じ単語表現を持つ複数の概念が存在する．したがって一般的な概念辞書である EDR を用いた場合も，一つの単語表現に対して複数の概念識別子が得られるため曖昧性が生じる．もう一つの原因は，一発話から得られる自立語数の少なさである．表 4 に

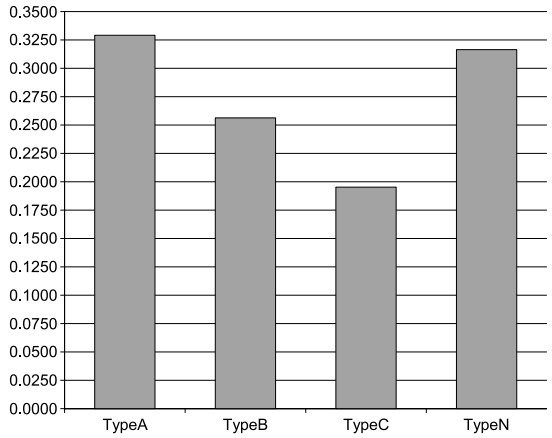

図 3: 話題・発話ベクトル間の発話タイプ別類似度傾向

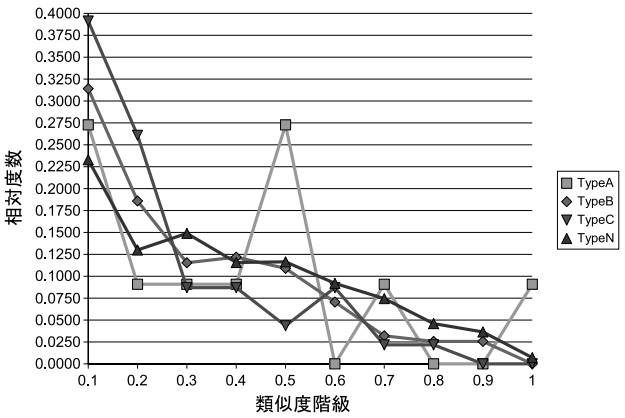

図 4: 発話タイプ別類似度相対度数分布

各コーパスの有効発話数と総抽出自立語数，一発話あたりの平均自立語数を示す．表から分かるとおり，一発話から得られる自立語は全体平均で 2.5 単語と少ない．

表 4: 有効発話自立語数

	有効発話数	抽出自立語数	平均自立語数
対話 1	413	860	2.1
対話 2	557	1281	2.3
対話 3	562	1654	2.9
合計	1532	3795	2.5

5.2 話題転換点の分類に有効な属性の分析

話題転換のリンクタイプによる影響を調べるため，発話ベクトルを 3. で示したリンクタイプごとに分離し，それぞれで話題・発話ベクトル間の類似度を計算した．

リンクタイプ Ex と Rex には基準概念より抽象的な概念と具体的な概念の両方を含んでいる．抽象性の違いによる類似度特徴の差異を見るため，リンクタイプ Ex，Rex をもつ概念は抽象度を用いて基準概念より抽象的な概念の ExA，RexA と具体的な概念の ExR，RexR に分離する．

EDR 概念体系辞書に記述されている概念体系は，最上位概念から末端概念までの階層の深さが一定ではないため，最上位概念からの距離のみではその概念の抽象度を設定できない．また多重継承が認められているため，最上位概念から対象概念に至るまでの経路が複数存在し得る．そこであるノードから最上位概念までと末端概念までの距離の比率と，そのノードをルートとするの部分木が概念木全体に対して占める割合を用いて係数を定義し，最上位概念から対象概念までの中間ノードがもつ係数の平均値で概念の抽象度を定義する．ある概念 n の係数は以下の式で定義する．

$$Rl_n = 1 - \frac{Ll_n}{Ll_n + Lr_n} \times \frac{N_n}{Na} \tag{8}$$

ここで

Ll_n : 概念 n の末端概念までの最長距離
Lr_n : 概念 n の最上位概念までの最短距離
N_n : 概念 n をルートとする部分木の全ノード数
Na : 概念木全体のノード数

である．したがって最上位概念の Rl_{root} は 0，末端概念の Rl_{leaf} は 1 である．

最上位概念からある概念 n までの経路が m 通りあったとすると，抽象度は以下のように定義される．

$$A_n = \frac{1}{m} \sum_{k=1}^{m} \frac{\sum Rl_k}{path_k} \tag{9}$$

ここで

m : 最上位概念から概念 n に到達可能な経路数
Rl_k : 経路 k での最上位概念から概念 n までの中間ノードの Rl
$path_k$: 経路 k での最上位概念から概念 n までの距離

である．最終的に抽象度は 0（最も抽象的）から 1（最も具体的）の間の値をとる．

こうして発話ベクトルをリンクタイプと抽象度を利用して ExA，ExR，Ex2，RexA，RexR，Bn，Cn の七つに分離した．なお ExA と RexA はその起点となる概念から最上位概念までの経路上にある各ノードに対し距離 1 以下の位置にある概念のみを残した．分離された七つの発話ベクトルとそれから構成される話題ベクトル間の類似度を発話ごとに求め，以下の分析を行った．

分析 1

TypeA，B，C の発話のみを取り出し，関連性のある話題転換（TypaA+B）と関連性のない話題転換（TypeC）を分類する上で重要となる属性を調べた．TypeA+B25 発話，TypeC25 発話，計 50 発話をランダムにサンプリングしたものを 1 セットとして 100 セットの訓練データを作成し，C4.5[3] で決定木を生成させ，情報利得比の最も高い最上位の分岐節点で用いられている属性を調べた．

分析 2

話題転換点と非転換点を分離する上で重要になってくる属性を調べた．全有効発話を話題転換点（TypeA+B+C）と非転換点（TypeN）に分け，各 50 発話，計 100 発話をランダムにサンプリングしたものを 1 セットとして 100 セットの訓練データを作成し，分析 1 と同様に C4.5 で決定木を生成させ最上位の分岐節点で用いられている属性を調べた．

結果

分析 1 および 2 の結果を表 5 に示す．表より，分析 1 では

表 5: 最上位の分岐節点で用いられる属性の比率

属性	分析 1	分析 2
ExA	19.57%	3.45%
ExR	7.61%	5.57%
Ex2	10.87%	1.15%
RexA	32.64%	3.45%
RexR	1.09%	14.94%
Bn	10.87%	64.37%
Cn	17.39%	6.90%

ExA や RexA が最上位の分岐節点で用いられることが多いのがわかる．この結果から，話題転換点の直前の話題との関連性の有無による分類には，発話中の語の概念と関連のある概念群の中で、発話中にある語の概念よりも抽象的な概念間での繋がりが重要であると考えられる．一方，分析 2 では RexR や Bn の比率が高くなっていることから，話題転換点・非転換点の分類には，発話中の語の概念と比較して同等かより低い抽象度の概念間における繋がりの強さが影響していると思われる．

6. 結論

発話・話題間の概念的関連性を計算するシステムを構築し，拡張概念を用いた話題転換点の分析を行った．分析の結果，話題転換点での発話・話題ベクトルの類似度には統計的な傾向がみられた．しかしながら話題転換点および転換タイプを発話が行われた時点で判別する指標としては不十分であった．これは単語と概念表現の対応の曖昧性や一発話から得られる情報の少なさ，および拡張概念であっても人間の持つ概念的関連性の知識を完全には表現できないために省略されている語を介しての関連性を適切に捉えられていないことが原因と考えられる．

また，話題転換時の関連性の有無による分類には，発話から得られる概念よりも抽象的な概念間での繋がりの強さが影響していることを示した．話題転換点と非転換点の分類においては，発話から得られる概念よりも具体的な概念での繋がりの強さが影響していることを示した．

今後は，発話に現れない概念間の関連を適切に判定可能な概念構造および発話や話題の概念を的確に表現する手法について検討を進める予定である．

参考文献

[1] 福地肇. 談話の構造. 大修館書店.

[2] 高梨克也, 井佐原均. 三者会話データの収録方法及び分析枠組みの概要. 言語処理学会第 8 回年次大会発表論文集, pp. 116–119, 2002.

[3] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993. (古川康一 監訳 (1995). 『AI によるデータ解析』, トッパン).