被放置発言からの物語マイニング

山口広樹，大澤幸生，西原陽子
東京大学大学院工学系研究科
〒113-0032 東京都文京区本郷 7-3-1 工 8-514

E-mail：gutsu@panda.sys.t.u-tokyo.ac.jp
{ohsawa, nishihara}@sys.t.u-tokyo.ac.jp

Abstract: 第 29 回ことば工学研究会で「例外発言」として扱ったものを，本論ではより具体化し「被放置発言」として扱い，そこから発言の背後に潜む発言者の物語をマイニングする手法を紹介する．加えてアンケートによる主観評価を用いた「被放置発言」の考察を，例を挙げて説明する．

1. はじめに

会話によるコミュニケーションは，社会学[井上 2000][長谷 09]や認知科学[坊農 04]，さらに工学[清田 03] [Alvarez 02]の研究分野としても様々な観点から扱われており，コミュニケーション研究において重要であることは論を持たない．そこで本論では，“果たして我々の日常会話は成功しているのかどうか”という問いに関して，会話解析を通して検証を試みる．さらに，“理想的な会話とはなにか”というおよそ正解が見えるとは考えられない問いにまで言及を試みる．

本論では，“会話内で周囲の会話参加者から放置されている発言”と定義する被放置発言に注目することで会話を検証する．一般に，放置されてしまう発言が多く存在する会話は，良い会話ではないと考えられる．ここで，「全ての発言が注目され，各々についてよく検討されている会話」を良い会話であると仮定すると，被放置発言が全く抽出されない会話こそが理想的な会話であり，被放置発言は否定的に捉えられるべきであるといえる．

しかしながらその一方で，「対等で平等な個人による相互理解」が理想的コミュニケーションではなく「“分からない”からこそコミュニケーションは楽しい」という観点も存在する[長谷 09]．この観点では，“分からない”ことをコミュニケーションによって理解しようと試みることでコミュニケーションを楽しむことができる．本論では，会話解析を通して，こうした観点も含め，被放置発言の抽出とそこに含まれる発言者の内包する物語をマイニングすることを目的とする．

また，本論の解析手法はテキストマイニング技術を応用したものであり，その基礎技術としてベクトル空間法[Salton 75]を用いる．ベクトル空間法は，文書要約[Hori 04][住田 95]や情報検索[Salton 83]など幅広く応用されており，シンプルかつ汎用性の高い手法であるため，本論でもこれを用いて解析を行う．また本手法は，話題構造分析[戸田 07]等の処理はせず，会話内で使用されている言葉のみを計算対象として解析するというシンプルな手法であるという特徴を持つ．本論ではオンラインチャット実験により評価する．

2. 被放置発言とは

前述のように，本論では被放置発言を“会話内で周囲の会話参加者から放置されている発言”と定義する．ここでは発言に含まれる単語のみを計算対象とし，各発言とその前後の会話との類似度を計算し，その類似度が低い発言を被放置発言として抽出する．

3. 被放置発言の抽出手法

まず，対象テキストデータを形態素解析[茶筌]し，そこで得られた単語群を(1)式の単語ベクトルに変換し，同様に各発言における出現単語をベクトルで表したものを(2)式の発言ベクトルで表す．ここで，w_k は k 番目に出現した単語の出現頻度であり，T は転置ベクトルを示す．

単語ベクトル　$W=^{T}(w_1, w_2, \cdots, w_n)$　(1)

発言ベクトル　$U=^{T}(w_1, w_2, \cdots, w_n)$　(2)

次に，生成した単語ベクトルと発言ベクトルを基に，発言と発言前後のコンテクストとの関連値を(3)式で算出する．本研究でのコンテクストの定義は文章の前後の脈絡とする．(3)式で，i は注目発言の発言番号，$W_{beforei}$ は注目発言前までの発言ベクトルの総和，$W_{after\cdot i}$ は注目発言後の発言ベクトルの総和，w は計算範囲窓を示す．

$$I_{beforei}=W_{before\cdot i}^{T}\cdot U \quad \left(W_{beforei}^{T}=\sum_{i-w\le k<i}U_k\right)$$
$$I_{after\cdot i}=U^{T}\cdot W_{after\cdot i} \quad \left(W_{after\cdot i}^{T}=\sum_{i<k\le i+w}U_k\right) \qquad (3)$$

(3)式で求めた発言ベクトルと前後のコンテクストとの関連値から発言性質を推定する．(4)式に示す条件式を満たす発言，つまり前後のコンテクストとの関連値が共に低いものを被放置発言として抽出する．ここで α,β は，それぞれ before 類似度，after 類似度の閾値であり，ユーザーが設定可能な変数ある．

$$I_{before\cdot i}<\alpha \cap I_{after\cdot i}<\beta \qquad (4)$$

4. 被放置発言の評価

3.で示した手法によって被放置発言を抽出した後，抽出した被放置発言の重要度を評価する．抽出される被放置発言は，本論の目的とする“言外に物語を含む発言”と“特に物語を含まない発言”が混在する．このため，目的に合致する発言であるか否かを評価する必要がある．

ここでは，選択平均情報量という新たな情報量を定義し，それを用いて被放置発言の評価を試みる．選択平均情報量は(5)式で定義される．

$$E_{\mathbf{AU}_k}=\sum_{t_i\in\mathbf{U}_k}p_{\mathbf{U}_k t_i}I_{\mathbf{A}t_i} \qquad (5)$$
$$\left(I_{\mathbf{A}t_i}=-\log p_{\mathbf{A}t_i}\right)$$

但し，
- t_i : i番目の単語の出現頻度
- $\mathbf{U}_k$: 注目発言ベクトル
- $p_{\mathbf{U}_k t_i}$: $\mathbf{U}_k$における単語 t_iの生起確率
- $p_{\mathbf{A}t_i}$: 集合$\mathbf{A}$における単語 t_iの生起確率

選択平均情報量は，注目発言が発言群に与えるインパクト値を表す．ここでは類似度と同様に，(6)式のように before 情報量と after 情報量を定義する．但し，計算範囲窓 w は類似度算出時と同じ値を用いる．

$$E_{\mathbf{W}_{beforek}\mathbf{U}_k}=\sum_{t_i\in\mathbf{U}_k}p_{\mathbf{U}_k t_i}I_{\mathbf{W}_{beforek}t_i}$$
$$E_{\mathbf{W}_{after\cdot k}\mathbf{U}_k}=\sum_{t_i\in\mathbf{U}_k}p_{\mathbf{U}_k t_i}I_{\mathbf{W}_{beforek}t_i} \qquad (6)$$

(6)式の二つの情報量はそれぞれ，注目発言の前までの会話に与えたインパクトの程度と，注目発言の後に与えたインパクトの程度を表している．算出した二つの情報量を基に被放置発言の性質を評価するために，(7)式のように評価関数を定義する．

$$\overline{E}=E_{\mathbf{W}_{beforek}\mathbf{U}_k}-E_{\mathbf{W}_{after\cdot k}\mathbf{U}_k} \qquad (7)$$

(7)式によって算出された評価値によって，表 1 に示すような発言の性質を評価できると考えられる．

表 1　評価値の性質

評価値 $\overline{E}$ 大	コンテクストに多少影響を与えており，発言者の主張・物語が含まれる可能性が大きい被放置発言
評価値 $\overline{E}$ 小	コンテクストに影響を与えておらず，発言者の主張・物語が含まれる可能性が小さい被放置発言

表 1 のように，評価値が大きい被放置発言に注目することで，会話内に埋没している物語をマイニングすることができると期待できる.

5. 結果出力・システム利用法

算出した評価値でランク付し，ランク，発言番号，発言者，発言内容，評価値をランク順に出力され，重要であれと思われる発言が注目されやすいように設計されている.

出力された解析結果から被放置発言の発言者の持つ物語を推測し，より具体的なテーマを設定した上で，追加インタビューやディスカッションを行う. これによって，先の会話内で話題として掘り下げられなかったテーマに注目しながら，より具体的・実践的な議論へと発展していくと期待できる.

6. オンラインチャット実験

6.1. 実験概要

本研究では，問題解決型のテーマを与えてそのオンラインチャット実験を行った. 日常会話でテーマとなるような簡単な話題をテーマとして与え，その会話ログを対象データとして解析し，提案手法の評価実験とする.

被験者は，同じ研究室に所属している大学院生で３～５人で一グループを作り，それぞれ数回オンラインチャットを行った. 実験の 1 回目は「研究室生活に関して」という比較的制約の弱いテーマで自由にチャットを行い，2 回目以降は，解析結果をフィードバックした後に被験者本人がテーマを決定するとうプロセスで行った. これにより，徐々に話題が掘り下げられていく形で実験が進んでいくようになる.

また，抽出した被放置発言の妥当性と重要度を評価するために，被験者に事後アンケートを行った. 本アンケートでは，「話題貢献度」と「思い入れ」の 2 項目について 5 段階評価で回答を得た. 各項目の定義を次に挙げる. ここでは，敢えて“どちらともいえない”という項目を設けることで，人間の感性の曖昧性を吸収できるように設定した.

【話題貢献度】

評価すべき発言が会話内で話題として取り上げられているかどうかの程度

【思い入れ】

評価すべき発言内と言外（言いたくても言えなかったこと，敢えて言及する必要のなかったこと等）に含まれる思いの強さ

ここで，各項目の評価基準は表 2 のように設定した.

表 2 アンケート評価基準

話題貢献度	1	放置された
	2	ほとんど貢献なし
	3	どちらともいえない.
	4	多少話題に取り入れられた
	5	話題に取り入れられた
思い入れ	1	思い入れはない
	2	大して思い入れはない
	3	どちらとも言えない. 分からない
	4	多少思い入れがある
	5	思い入れがある

「話題貢献度」によって抽出した被放置発言の妥当性が，「思い入れ」によって被放置発言に含まれる物語の有無と関心の強さが評価できると考えられる. また，「思い入れ」の評価にあたって，評価者本人以外の発言に関しては，発言者の「思い入れ」を推測して評価するように指示した.

本アンケートでは，抽出した被放置発言と，ランダムに選ばれたその他の発言を無作為に評価用発言として提示される. 被験者は提示された各発言に対して，発言前後の会話を参照しながら評価を行うという形式で，評価専用のHTML 形式ファイルを利用することで，容易に評価が行える環境で実施した.

6.2. 結果と考察

オンラインチャット実験は全 7 回実施され，事後アンケートでは，延べ 20 名，全 77 発言，253 回答（内，被放置発言 112，その他発言 141）が得られた．抽出した被放置発言のうち，物語をマイニングするトリガーとなった発言の一部を表 3 に示す．なお，ここでは発言者名，発言番号は省略し，ランクは各発言の該当するチャット内でのランクを示す．

また，実験は 4 グループで各々行ったものであり，2 回目以降のチャットテーマはそれぞれ異なる．フィードバックとして被放置発言を提示することによって，会話内ではあまり注目されなかったにも関わらず，次の会話のテーマに選ばれることになった発言もあり，メッセージ性の高い被放置発言が存在することが実験により示された．より具体的な解析内容に関しては後述する．

表 3 解析実験結果（一部抜粋）

実験番号	ランク	発言内容	評価値
A1	3	どうりでまだ残ってる	0.028
A2	3	その機能あったらすごいw	0.44
B1	4	代わりにMr. Yでお許しを	0.021
B2	4	ただ，1こ下にすでに17という化け物がいます	0.0082
B2	6	寄付金はしっかり払ってくれます	-0.066
B3	12	やっぱり中庭かな	-0.24
C1	1	T（電子楽器）でクリスマスソングか	0.29
D1	2	ところで，研究室でバスケとかしたいね！	0.69
D2	4	T（電子楽器）featuring Y	0.29
D2	5	あの指向性スピーカー使いませんか？w	0.24

表3に示した被放置発言から物語を掘り起こすためには，前後のコンテクストを参照する必要があり，それによって初めて被放置発言がその潜在力を発揮することになる．本論では例として，表3から3発言に注目して以下に前後の会話ログとともにその解析結果を挙げる．

- 代わりにMr. Yでお許しを

----------会話ログ----------
G > 火を使うには申請が必要だとか
Z > 申請なしで報知器なったらみんなでダッシュだな
Z＞BとSさんを残そう
Z> じゃあFくんも残ろうw
F＞代わりにMr. Yでお許しを
Z＞Y「私めが残るのでございますか」
-----会話ログここまで-----

会話ログを参照すると，当該発言は“完全に放置されてしまった発言”ではなく，コンテクストに影響を与えているものの同じ単語は使われなかったために被放置発言として抽出されてしまったと考えられる．よって，被放置発言の定義に沿わない発言と言えるが，この発言には物語が含まれていると推測される．

後に追加インタビューを行ったところ，この発言に登場するY氏に対して「潜在能力が高く期待できる人材だが，それを未だ発揮できていないように思われ，今後に期待している．」という回答を得られ，Y氏への関心の高さが感じられた．

以上のように本発言は，実際には放置されていないがコンテクストに影響を与えており，物語を含んでいる発言が被放置発言として抽出された例となる．また，ここで交された会話自体は他愛のない日常会話であるが，何気ない会話の中に所属するコミュニティ固有の関心事が表れており，会話参加者が共通に持っている認識が被放置発言として抽出されていることが分かった．

- T（電子楽器）でクリスマスソングか

----------会話ログ----------
H＞みんな何かしら一発芸をやるとか
T > テノリオンでクリスマスソングか
H > お，T君の一発芸決まったね！ｗ
-----会話ログここまで-----

電子楽器Tは，被験者らが所属する研究室で一時期話題となっていたものであり，“クリスマス会で利用したらおもしろいのではないか”という発言．電子楽器Tの有効利用方法を模索しているという物語が浮かび上がってくる．

- あの指向性スピーカー使いませんか？ｗ

----------会話ログ----------
M >それいいねぇ．Y君のアニソン祭り．
S >Tにあわせて歌うw
M >ステージも用意してあげなきゃ．マイクはマイマイクを持ってきてもらおう．
S >あの指向性スピーカー使いませんか？ｗ
Y>誰も言ってないのに持ってくるY君…
Y>K先生にYビーム
M >おぉー！それはいい．そして山田君は伝説となる！
-----会話ログここまで-----

この発言に関しても，会話内容に沿っていないような発言ではないがコミュニティを特徴付ける物語がマイニングできる発言が抽出された．“研究室で所持している指向性スピーカーと前述の電子楽器Tを同時に有効利用できる機会をつくることができるのでは”という期待が見受けられる発言となる．

以上のような解析結果から，暗に受け入れられた概念は，明示的に発話されないため，コンテクストに影響を与えているにも関わらず被放置発言として抽出されることが分かった．

実際には放置されていないものの被放置発言として抽出されている発言も多く存在する．しかしながら，これらの発言には物語が内包されていることが多いことも本実験により明らかにされた．この理由として，「コミュニティ内で共通の意識として認識されていることは，会話内で明示的に言葉として表れることが少ない」ために生じる現象であると考えることができる．

また，コミュニティの特徴を掴むトリガーとなる発言を抽出できるため，コミュニティの非所属者であっても，そのコミュニティ固有の関心事や問題意識を理解する助けとなることが実験により明らかになった．

6.3. 評価アンケート結果

チャット実験のアンケート結果を図1に示す．図1では，円の大きさが回答数を表しており，「話題貢献度」と「思い入れ」に正の相関が見られる．これより，「思い入れ」が強い発言ほど話題に貢献しており，全体として良い会話であると考えられる．

次に円グラフに注目する．円グラフは，得られた回答のうち被放置発言が占める割合を示す．ここで，放置されてしまった発言は「話題貢献度」が低くなるため，被放置発言の抽出精度が高ければグラフの左側が黒に染まるという結果が理想である．アンケート結果から，「話題貢献度」に注目すると，精度が高いとは言えず，t検定を行った結果，本実験結果において被放置発言とその他発言との「話題貢献度」の平均値に有意差があるとは言えなかった．従って，被放置発言の抽出手法に関しては，今後解決すべき課題の一つとなる．

一方「思い入れ」に関しては，図1より，被放置発言の「思い入れ」は比較的強いように見てとれる．また，t検定を行ったところ，被放置発言

とその他発言との「思い入れ」の平均値に有意差が認められた（P<0.05）.

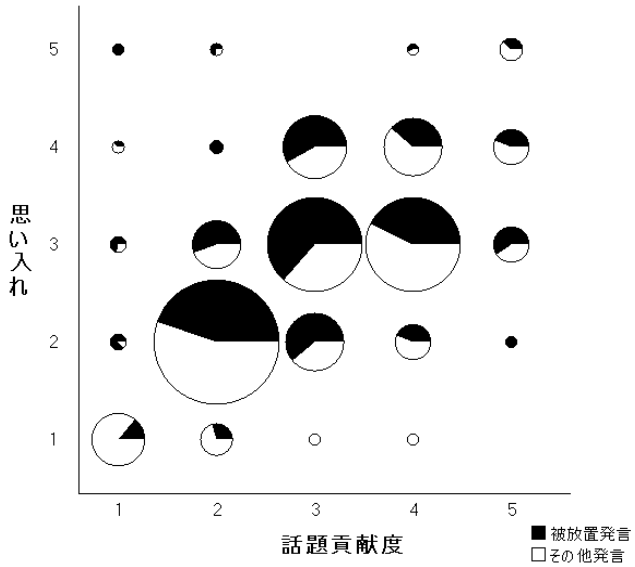

図1　アンケート結果可視化グラフ

7. おわりに

本論では，被放置発言に注目して，被放置発言を抽出する手法とその評価手法を用いて，会話に内包される物語をマイニングする手法を提案した．ここでは，会話内で使用されている言葉のみを入力データとして，各発言とその前後コンテクストとの関係を評価することで被放置発言を抽出し，さらに選択平均情報量を用いて被放置発言の性質を計算によって評価した．

また，提案手法の評価のためにオンラインチャットを用いた実験を行った．評価実験は，同コミュニティに所属する被験者を対象に行い，被験者の属するコミュニティ特有の関心事や問題意識を抽出することができた．コミュニティに内在する物語をマイニングすることで，コミュニティ所属者は所属しているコミュニティに関して再認識することができ，且つ，コミュニティの非所属者がコミュニティを理解するための役に立つことが分かった．被放置発言に注目して解析を進めたにも関わらず放置されていない発言も多く抽出され，抽出エラーが多く今後の課題であるいえるが，そこから物語をマイニングすることができ，結果的には目的を一部達成できたといえる．また，被放置発言の評価に利用した選択平均情報量に関しての検証も今後の課題の一つといえる．

謝辞

実験にご協力していただいた東京大学大学院工学系研究科機械工学専攻の金子・山崎研究室の皆様に感謝いたします．

参考文献

[坊農 04] 坊農 真弓，鈴木 紀子，片桐 恭弘．多人数会話における参与構造分析 ：―インタラクション行動から興味対象を抽出する．*認知科学*, **11**, 214-227, 2004.

[井上 00] 井上隆二，山下富美代，図解雑学社会心理学（図解雑学シリーズ），ナツメ社, 2000.

[長谷 09]長谷正人，奥村隆，コミュニケーションの社会学，有斐閣アルマ, 2009.

[清田 03] 清田陽司，黒橋禎夫，木戸冬子，大規模テキスト知識ベースに基づく自動質問応答ーダイアログナビー，自然言語処理, Vol.10, No.4, pp145-175, 2003.

[Alvarez 02] Rosío Alvarez, Discourse Analysis of Requirements and Knowledge Elicitation Interviews, Proceedings of the 35th Hawaii International Conference on System Sciences, 2002.

[Salton 75] G. Salton, A. Wang, C. S. Yang, A Vector Model for Automatic Indexing, Communication of the ACM, Vol.18, No.11, pp.613-620, 1975.

[Hori 04]Chiori Hori and Sadaoki FURUI, Speech Summarization: An Approach through Word Extraction and a Method for Evaluation, IEICE TRANS. INF. & SYST., Vol.E87-D, No.1, 2004.

[住田 95] 住田一男，知野哲郎，小野顕司，三池誠司，文書構造解析に基づく自動抄録生成と検索提示機能としての評価，電子情報通信学会論文誌 D Vol.J78-D2, No.3, pp.511-519, 1995.

[Salton 83]G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[戸田 07]戸田浩之ほか，グラフ分析を利用した文書集合からの話題構造マイニング，電子情報通信学会論文誌 D Vol.J90-D , No.2, pp.292-310, 2007.

[茶筌] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸，日本語形態素解析システム『茶筌』version2.2.1 使用説明書, 2000.