例外発言からの価値マイニング

A Value Mining from the Exceptional Utterances

*山口広樹，西原陽子，大澤幸生
東京大学大学院工学系研究科
Hiroki Yamaguchi, Yoko Nishinara, and Yukio Ohsawa
School of Engineering, The University of Tokyo
*Email: gutsu@panda.q.t.u-tokyo.ac.jp

キーワード:テキストマイニング，選択平均情報量，例外発言
Keywords: Text Mining, Average Amount of Choice Information, and Exceptional Utterance

1. はじめに

飛躍的に発展した近年の科学技術では膨大な量の情報が扱われるようになり，その中から有用な情報を見つけ出す技術としてデータマイニングの研究が進められてきた．従来のデータマイニングでは稀に起こる事象はノイズとして除去されてしまうことが多く，潜在的価値を有している情報もノイズとして除去されてしまっている可能性が高い．稀な事象にこそチャンスが潜んでいるとして単語の共起度に注目して情報の可視化を行う研究としてチャンス発見学[1]が大澤らによって提唱されている．本研究は，チャンス発見学の次世代研究となるバリューセンシングという新たな枠組みの研究の一つとなる．本研究の目的は，例外発言を抽出してその価値を再検討することで新たな価値を創造することである．

例外に注目した関連研究としては，情報量を用いてデータベースから例外的知識を抽出するもの[2]や例外から得られる情報を基にプログラムの解析やテストをするもの[3]などがあるが，これらはデータベースやプログラムなど構造的なデータを対象としている．本研究では，構造は無秩序で流動的であるが率直な意見・主張が含まれていると考えられる自然言語による会話を解析対象とする．

自然言語解析では，用いられている単語群をベクトルで表して処理するベクトル空間法[4]が情報検索などに広く利用されている．ベクトル空間法はシンプルであるが故に汎用性が高いと考えられる．構造が無秩序である会話を扱うために本研究ではこれを採用し，二つのテキスト間の類似度をベクトルの内積によって求める．前後のコンテクストとの類似度が低い発言，つまり会話の流れに乗っていない発言を例外発言として抽出する．次に，抽出した例外発言を評価するために選択平均情報量という情報量を定義して用いる．選択平均情報量とは，ある全集合の部分集合の持つ全集合における情報量の期待値である．選択平均情報量を基に，会話コンテクストにおける発言性質を推定することができる．ここで，発言の性質とは，発言前後のコンテクストに与えている（または与えられた）影響度から推定できるものとし，3.で定義する．

また，抽出した例外発言から新価値を創造していくプロセスは，人と計算機の各々の処理を交互に繰り返すことによって相乗効果をもたらす二重らせんモデル[1]を採用する．

2. 提案システム概要

本研究では，例外発言を抽出・評価してその価値をマイニングしていく.図１に提案システムにおけるテキスト処理部のフローチャートを示す.

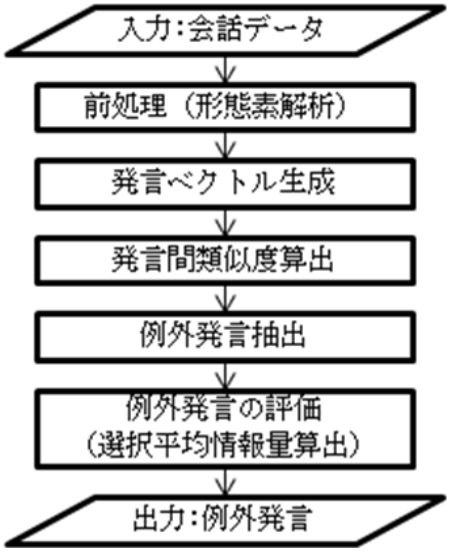

図１ テキスト処理部フローチャート

3. 例外発言の抽出―発言間類似度―

前処理として会話テキストを形態素解析し，その結果得られる単語群を発言ベクトルとして(1)式で表す．但し，t_k は k 番目の単語の出現頻度，n は総単語種類数である．

$$\mathbf{U}_k = (t_0, t_1, \cdots, t_n) \qquad (1)$$

二つの発言ベクトルの内積を類似度とするが，ここでは注目発言とその前後のコンテクストとの類似度を求めるため，(2)式で before 類似度と after 類似度を定義する．

$$I_{before\cdot k} = \mathbf{W}_{before\cdot k}^{T} \cdot \mathbf{U}_k$$
$$I_{after\cdot k} = \mathbf{U}_k \cdot \mathbf{W}_{after\cdot k}^{T} \quad (2)$$
$$\text{但し,} \begin{pmatrix} \mathbf{W}_{before\cdot k} = \sum_{i<n} \mathbf{U}_i \\ \mathbf{W}_{after\cdot k} = \sum_{i>k} \mathbf{U}_i \end{pmatrix}$$

(2)式の before 類似度と after 類似度が共に小さい発言を例外発言として抽出する．本稿では経験的に閾値を設けて抽出を行う．

4. 例外発言の評価―選択平均情報量―

会話コンテクストにおける各発言の位置付を評価するために，発言群内における一発言の情報量を選択平均情報量として(1)式で定義する．(1)式は，集合 A における一発言 $\mathbf{U}_k$ の平均情報量を表す．

$$E_{\mathbf{AU}_k} = \sum_{t_i \in \mathbf{U}_k} p_{\mathbf{U}_k t_i} I_{\mathbf{A}t_i}$$
$$I_{\mathbf{A}t_i} = -\log p_{\mathbf{A}t_i} \quad (3)$$
$$\text{但し,} \begin{Bmatrix} t_i : i\text{番目の単語の出現頻度} \\ \mathbf{U}_k : \text{注目発言ベクトル} \\ p_{\mathbf{U}_k t_i} : \mathbf{U}_k\text{における単語}\, t_i\text{の生起確率} \\ p_{\mathbf{A}t_i} : \text{集合}\mathbf{A}\text{における単語}\, t_i\text{の生起確率} \end{Bmatrix}$$

(3)式の集合 A を注目発言の前後(注目発言を含む)の発言群として(4)式のように算出し，(5)式で，得られた二つの選択平均情報量の差を評価することで抽出した例外発言の発言性質を推定する．

$$E_{\mathbf{W}_{before k}\mathbf{U}_k} = \sum_{t_i \in \mathbf{U}_k} p_{\mathbf{U}_k t_i} I_{\mathbf{W}_{before k} t_i}$$
$$E_{\mathbf{W}_{after\cdot k}\mathbf{U}_k} = \sum_{t_i \in \mathbf{U}_k} p_{\mathbf{U}_k t_i} I_{\mathbf{W}_{before k} t_i} \quad (4)$$
$$\text{但し,} \begin{pmatrix} \mathbf{W}_{before\cdot k} = \sum_{i \le n} \mathbf{U}_i \\ \mathbf{W}_{after\cdot k} = \sum_{i \ge k} \mathbf{U}_i \end{pmatrix}$$

$$\overline{E} = E_{\mathbf{W}_{before k}\mathbf{U}_k} - E_{\mathbf{W}_{after\cdot k}\mathbf{U}_k} \quad (5)$$

上式で前後の会話コンテクストにおける注目発言の情報量を算出でき，これが選択平均情報量となる．選択平均情報量が大きいと，選択した部分集合の持つ，全集合内での情報量が多くなる．ここでは全集合が注目発言前後の発言群，選択した部分集合が注目発言となるため、選択平均情報量は注目発言が前後のコンテクストにどれほど影響を与えているかを示し，前後の会話コンテクストとのズレが大きいほど値が大きくなる．また，(5)式の値が大きいほど周囲に受け入れられた発言であり，理想的な例外発言であると言える．

5. 提案システムとその利用法

本研究では，会話テキストを入力として，形態素解析・発言ベクトル生成・類似度算出・選択平均情報量算出という処理手順で例外発言を抽出，評価するという手法を提案しているが，こうして得られた例外発言から価値をマイニングすることが主な目的となる．

本手法を用いて新価値を創造していくプロセスは二重らせんモデル[1]を採用する．ここでは，会話データから抽出した例外発言をフィードバックして再度会話する，というように繰り返し行うことで例外発言から価値マイニングする．

6. おわりに

本稿では，潜在的価値が埋もれてしまっていると考えられる例外発言えおテキスト処理によって抽出し，その価値を再検討するというプロセスを繰り返すことによって例外発言からの価値マイニングを行う手法を提案した．

今後は，ディスカッションの解析などの実験をし，本手法の有効性を検証していく．実験では，本手法を用いてマイニングしたアイデアをアンケート調査によって評価し，本手法によって有効なアイデアが得られることを確認する．

参考文献

[1] 大澤幸生，チャンス発見の情報技術，東京電機大学出版局, 2003..

[2] 鈴木英之進，志村正道，情報理論的手法を用いたデータベースからの例外的知識の発見，

[3] 人工知能学会誌, Vol.12,No.2, pp.305-312, 1997.

[4] Saurabh Sinha, Mary J. Harrold, Analysis and Testing of Programs with Exception Handling Constructs, IEEE TRANSSACTION ON SOFTWARE ENGINEERING, Vol.26, No.9, pp.849-871, 2000.

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