2S-3

チャット文章の利用目的に着目した分類方法の提案

原田修平† 丸山広† 高嶋章雄† 中村太一†
東京工科大学大学院†

1.まえがき

プロジェクトマネジメント教育にオンライングループワーク環境でのロールプレイ演習が用いられている．演習者の行動を分析して，指示をすることで教育効果を高められる[1]．特に，分析結果に基づいたフィードバックを実時間で行うことが求められている．しかし，演習者のチャット文章を実時間で分析することは難しい．チャット文章は，相槌や返事が多く．その特徴を捉えにくく単純な機械分類では分析しにくい．

本研究では，文章分類方法のキーワードベクトル法でチャット文章を分類する．ただし，短い文章の多いチャットの文章を分類するには，連続する文章の結束性を利用して，チャット文章を集合に分割し，分割された集合に付与されたカテゴリを，キーワードベクトル法で分類できなかった文章に付与することで，機械分類の精度を向上させる．

以降，2 で対象とするチャットデータの性質，3 で関連研究，4 で解決方法，5 で提案手法を述べ 6 で提案手法の結果を示し，7 で提案手法の有効性を示す．

2.ロールプレイ演習で取得するチャット文書

本稿では，2009 年度に実施したロールプレイ演習での 32 グループのチャットの 2491 発言を分類する．学習者のロールプレイ演習の取り組み態度をモニタするため，表 1 のカテゴリに分類した．

表 1 発言を分類するカテゴリ

カテゴリ	発言の意味
R	演習で指定された役割の発言
L	課題の解き方や知識に関する発言
M	システムの使い方に関する質問
O	その他の雑談など

The proposal of a method for classifying chat-messages
†Shuhei HARADA †HiroshiMARUYAMA
†AkioTAKASHIMA †Taichi NAKAMURA; Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology

しかし，チャットの発言内容を表 1 のカテゴリに分類する際に以下の 2 つの問題がある．

(1) 相槌のような短い表現や似ている文章を機械分類で分類できない．
(2) 発言のカテゴリの分布に偏りがあるため，集合を分類することが難しい（表 2）．

表 2 カテゴリに出現する発言の分布

カテゴリ	R	L	M	O
分布率	81.7%	7.2%	5.7%	5.4%

チャットの文章には，「はい」や「なるほど」などの返事や相槌がある．このような短い文章はキーワードベクトル法で用いる特徴がなく機械分類できない[2]．また，相槌や返事の類似の表現は意味が違う場合がある．特にカテゴリ R に分類された発言の文章と L に分類された文章は，共にロールプレイの内容に関する文章であるため，その特徴が似ている．

3.関連研究

塩崎らは，製品の設計打合せの会話に含まれるひらめきやアイデアを再利用するため，会話文章からアイデアに関する話題を抽出し，それと関連する話題を発見する方法を提案している[3]．会話文にある短い文章では，文章の特徴を得ることが難しい．そのため，25 文章のまとまり（ウィンドウ）を用いた近似法とキーワードベクトル法を用いて，似ている文章を検索した．

これら方法は 50 文章以上のテキストデータを対象としており，チャットの文章のような短い文書を分類することできない．また，相槌などの短い表現を分類することも出来ない．

4.解決方法

2 で述べた 2 つの問題の解決方法を以下に示す．

(1) キーワードベクトルで分類できなかった相槌や特徴が見つからない文章を，それが含まれる集合のカテゴリに分類する．
(2) 分類の偏りがあるチャットの集合を分類

するため，分布の確率と分類確率の差分を用いて分類する

5.提案手法

チャットの文章には特徴量が多い発言と特徴量が少ない発言がある．そのため，塩崎らの手法と TextTiling と組み合わせることで，相槌のような短い文章が続いても，比べるウィンドウに含まれる特徴量を動的に変化させることで，集合で分割する[4]．集合を構成する文章をカテゴリに分類するため，以下の式を用いて分類確率 $P(d)$ を求めた．この確率が最も高かった確率に分類する．

$$P(d)=\frac{\sum p(dx_i \mid \text{カテゴリ}_i)}{\text{語の数}}-p(d \mid \text{カテゴリ}_i)$$

ただし， $p(dx_i \mid \text{カテゴリ}_i)$ は，語 dx_i がカテゴリ $_i$ に分類される確率であり， d は，集合 dx_i を含む文章集合である．

キーワードベクトルを用いた，チャットの文章の分類で，分類できなかった文章を，これらの文書集合のカテゴリ分類を利用して分類する．

6.結果

キーワードベクトル法の分類結果の再現率・適合率を表 3，文書集合の分類結果の再現率と適合率を表 4，提案手法で分類した結果を表 5 に示す．

表 3　キーワードベクトルの再現率・適合率

カテゴリ	R	L	M	O
再現率	98.6%	0.0%	2.7%	7.2%
適合率	83.7%	0.0%	100.0%	100.0%

表 4　話題集合の分類結果の再現率・適合率

カテゴリ	R	L	M	O
再現率	52.0%	64.2%	63.0%	29.4%
適合率	90.5%	15.0%	15.2%	28.4%

表 5　提案手法の再現率・適合率

カテゴリ	R	L	M	O
再現率	89.3%	11.3%	15.1%	19.0%
適合率	84.8%	12.1%	13.9%	43.9%

7.考察

表 3，表 5 の各手法の再現率・適合率を比較する．表 3 のキーワードベクトル法では全く分類できなかった L の再現率・適合率が，提案手法では再現率 11.3%・適合率 12.1%となっている．

他にも，M の再現率が 2.7%から 15.1%に増加し，O の再現率も 7.5%から 19.0%に増加した．しかし，M の適合率は 100%から 13.9%に，O の適合率も 100%から 43.9%に減少した．これは，M や O に分類された文章が増えたためである．

着目していた問題点である R と似ている L の文章や O の文章の分類の再現率が増加している．そして L は，適合率も増加している．このことから，R に分類されたキーワードベクトルのみの分類と比べて，分類できていると言える．

キーワードベクトル法のみの分類結果を実際に見たところ，「ごめんなさい」「おお」という短い文章や，「なるほど」「同意」という相槌の分類ができなかった．提案手法での分類では，これらの文章の 30.2%を正確に分類できた．

8.あとがき

ロールプレイ演習のチャット文章を用いて，フィードバックに必要な文章分類を自動化するため，文章の構造化と機械分類を用いて分類した．その結果，従来の機械分類だけでは分類できなかった文章の分類に成功した．そして，構造化の精度向上により，更に分類精度が向上することがわかった．

参考文献

[1] Taichi Nakamura,etc,:"Method for designing a role-play scenario based on UML and evaluation of educational effect", Proceedings of 9th Joint Conference on Knowledge-based Software Engeneering (JCKBSE'10) Kaunas, Lithuania, (2010, in print)

[2] Gerard Salton, etc,:"Term-weighting approaches in automatic text retrieval", Information Processing & Management Volume 24, Issue 5, 1988, Pages 513-523

[3] 塩崎敏也,他,:" 設計における談話の分析と構造化", 電子情報通信学会技術研究報告. AI, 人工知能と知識処理 97(631), 41-48, 1998-03-26

[4] Hearst, M. A.:TextTiling:"Segmenting Text into Multi-paragraph Subtopic Passages, Association for Computational Linguistics", Vol.23, No.1,pp.111-112(1997)

[5] Jason D. M.,etc,:"Tackling the Poor Assumptions of Naive Bayes Text Classiers",Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003