協調学習における韻律特徴を用いた発話タグ推定モデル

Estimating Utterance Tags using Prosodic Features in Collaborative Learning

林 佑樹[*1], 大佛 駿介[*2], 中野 有紀子[*2]
Yuki HAYASHI [*1], Shunsuke OSARAGI[*2], Yukiko I. NAKANO[*2]
[*1] 大阪府立大学
[*1]Osaka Prefecture University
[*2] 成蹊大学
[*2]Seikei University
Email: hayashi@kis.osakafu-u.ac.jp

あらまし：協調学習において，説明や質問といった参加者が発する発言の意図（以後，発話タグ）を推定できれば，学習時の議論状況をリアルタイムに評価できるようになる．本研究では，協調学習における既存の対話コーパスデータの各発言に発話タグを付与し，ピッチや話速，抑揚といった韻律特徴に基づき，機械学習により発話タグを推定するモデルを提案する．評価実験の結果，4 割強の精度で 6 種類の発話タグを正しく推定できることを確認した．
キーワード：協調学習，発話タグ，韻律特徴，音声会話

1. はじめに

協調学習は主に参加者の対話によって進展するため，テキスト・チャットなどで参加者によって各発言に付与される説明や質問，同意といった発話タグに基づき学習時の対話状況を分析する研究が行われている[(1,2,3)]．このような発話タグを自動的に推定することができれば，学習時の対話状況をリアルタイムに評価できるようになる．しかし，発話タグを自動推定する研究の多くは，テキスト対話を対象としており[(4)]，コミュニケーションとしてより自然な音声対話は対象とされてこなかった．音声認識を利用する手法も考えられるが，多人数会話の場合，特に誤認識が多く生じてしまうという問題がある．

本研究では発話におけるピッチや話速，抑揚といった韻律特徴に着目し，協調学習における参加者の発話タグを自動推定することを目的とする．協調学習に特徴的な発話タグを既存のコーパスに付与し，音声の韻律特徴に基づき，機械学習を用いて発話タグを推定するモデルを提案する．

2. 分析データ

2.1 発話タグの付与

本研究では，先行研究で構築した協調学習マルチモーダルコーパス[(5)]に収録されている，知識共有タイプの課題に取組む 10 グループ分（1 グループ 3 名）の会話音声データを利用する．

発話タグとして，一般的な会話における 42 種類の談話タグが定義された SWBD-DAMSL タグ[(6)]を参考にした．2 名のアノテータによる付加作業を通し，使用する必要がないタグや，新たに追加すべきタグを選定し，文法レベル 14 個，文意レベル 6 個の発話タグを付与することとした．タグ付けの妥当性を検証するために，1 グループのデータに対して 2 名が独立で発話タグを付与した結果，観察者間の一致度は十分であることを確認した（*Cohen's* κ=0.69）．

音声から書き起こされた 2559 個の各発言に発話タグを付与した結果，各発話タグのデータ数にはばらつきが見られた．そこで，議論状態を推定するために重要と考えられる「平叙文」,「疑問文」,「同意」，そしてデータ数が比較的多い「相槌」,「笑い」，これら以外を「その他」とした計 6 種類を本研究の推定対象として扱う．

2.2 韻律特徴データの付与

発話タグと音声との関係を分析するために，コーパスに記載されている発話開始／終了時間情報に基づき，発話の音声データ（wav 形式, 2512 ファイル）を音声切り出しソフトウェア WAVEFLT2 により抽出した．ここで，ソフトウェアが抽出できなかった非常に短い発話は除外している．

韻律データを取得するために，音声分析ソフトウェア Praat[(7)]を使用した．本研究では「ピッチ」,「話速」,「抑揚」,「インテンシティ」および「ポーズ」の韻律特徴に着目する．ピッチ（Hz）は，発話における 0.1 秒毎のデータを取得した．話速は Praat から算出されるシラブル数と発話時間から，1 秒当たりのシラブル数を求めた．抑揚はピッチの最大値から最小値を引いた値とし，インテンシティ（dB）は Praat スクリプトに基づき算出した．ポーズ（sec）は，前発話からの経過時間に応じて 8 種類のタグを設定した．ここでは，前発話から 0.5 秒刻みで 3 秒後までのポーズを表すタグを 6 種類，3 秒以上経過してから発言されたタグを 1 種類，また発話が他の発話区間と重複していることを表すタグを 1 種類付加した．

3. 発話タグ推定モデル

3.1 推定に利用する特徴

発話タグ推定モデルを構築するために，2.2 節で述べた音声韻律特徴に加えて，前発話タグ，ピッチの頻度情報，性別の 3 種類を考慮した．

前発話タグは，発話の直前になされた発話のタグ

を表す．相槌に関しては 1 つの発話に対して複数回なされる傾向が見られたため，1 つの発話に対して複数の相槌が打たれた場合，すべての相槌の前発話タグはその発話タグとした．

ピッチの頻度情報として，全発言のピッチ平均情報を k-means 法を用いて 10 個のクラスタに分類し，1 発話の 0.1 秒ごとのピッチのフレーム値が各クラスタにどの程度分類されるか求めた．ピッチの値が取得できなかったフレームの場合は，非取得用のクラスタを 1 個用意し，そこに分類した．

最終的に発話タグの推定に使用する属性は，ピッチ 3 種類（最大値，最小値，平均），インテンシティ 2 種類（最大値，最小値），抑揚，話速，発話時間，前発話タグ，ポーズ時間，ポーズタグ，性別，ピッチの頻度（11 種類）の計 23 種とした．

3.2 モデル作成に用いる学習アルゴリズムの検討

推定モデルの作成のために，機械学習ソフトウェア Weka[8]を使用した．発話数が 200 を超える発話タグ（平叙文，疑問文，相槌）に関しては，リサンプリングを行い，サンプル数の上限を 200 に設定した．本モデルをリアルタイムで利用する状況を想定し，正確に推定することが困難だと考えられる前発話タグの有無を考慮した 2 種類のモデルを構築することとした．

10-fold 交差検証法によるモデル評価を行い，結果が良好であった 5 種のアルゴリズムを比較検討した．表 1 に各アルゴリズムにおける発話タグ分類の F-measure の値を示す．リアルタイムでの使用も考慮すると SVM が最も高い精度となることが示された．モデルの推定精度を高めるために，精度を下げる属性を除外する属性選択を行った．結果，前発話タグを属性に含める場合は 16 種，含めない場合は 17 種の属性がモデル推定のために選択された．

表 1 学習アルゴリズムの比較結果

アルゴリズム	モデル 1 （前発話タグ有り）	モデル 2 （前発話タグ無し）
C4.5	0.421	0.354
NN Search	0.382	0.350
NB Tree	0.473	0.399
Random Forest	0.411	0.404
SVM	0.460	0.443

4. 発話タグ推定の評価

属性選択で得られた属性を用いて，前発言タグ情報を利用するモデル（モデル 1），利用しないモデル（モデル 2）の 2 種類の発言タグ推定モデルを作成した．分類器は SVM を使用し，評価には Leave one out 交差検証法を用いた．

表 3 に発言タグ推定の F-measure を示す．全体として，両モデルともに 4 割強の精度で発話タグを推定できており，ランダムな推定（1/6≒0.167）と比較して 2 倍を上回る精度となった．特に「相槌」や「笑い」は他と比べて高い値を示している．「疑問文」は相手に同意を求めるために，平叙文直後に現れる付加疑問文が多かったため，精度が落ちたと考えられる．「同意」については，機械学習にかけたサンプル数が少なかったことが原因として挙げられる．

今回は音声の韻律特徴に着目しているが，今後は音声認識結果や視線対象などの情報を統合したモデルに拡張することで精度を向上できる可能性がある．

表 2 発話タグ推定モデルの推定結果

発話タグ	モデル 1 （前発話タグ有り）	モデル 2 （前発話タグ無し）
平叙文	0.452	0.419
疑問文	0.313	0.358
同意	0.277	0.022
相槌	0.662	0.579
笑い	0.650	0.658
その他	0.426	0.140
平均	0.486	0.452

5. おわりに

本研究では，協調学習における個々の発話に対して発話タグを手動付与し，音声データから取得される韻律特徴に基づいて発話タグを推定するモデルを提案した．全体として，精度は 4 割強の精度となることを確認した．今後の課題として，韻律特徴以外の情報を統合して推定精度を高める手法を検討していく予定である．

謝辞

本研究は JSPS 科研費 25280076, 26870588 の助成による．

参考文献

(1) Inaba, A., and Okamoto, T.: “The Network Discussion Supporting System Embedded Computer Coordinator at the Distributed Places”, Educational technology research, Vol.18, Nos.1-2, pp.17–24 (1995).
(2) 小谷哲郎，関一也，松居辰則，岡本敏雄：“好意的発言影響度を取り入れた議論支援システムの開発”，人工知能学会論文誌，Vol.19，No.2，pp.95‐104 (2004).
(3) Kojiri, T., Yamaguchi, K., and Watanabe, T.: “Topic-tree Representation of Discussion Records in Collaborative Learning Process”, Journal of Information and Systems in Education, Vol.5, No.1, pp.29–37 (2006).
(4) 磯村直樹，鳥海不二夫，石井健一郎：“対話エージェント評価におけるタグ付与の自動化”，電子情報通信学会論文誌 A, Vol.J92-A, No.11, pp.795-805 (2009).
(5) 林佑樹，小川裕史，中野有紀子：“協調学習における非言語情報に基づく学習態度の可視化”，情報処理学会論文誌, Vol.55, No.1, pp.189-198 (2014).
(6) Jurafsky, D., Shriberg, L., and Biasca, D.: “Switchboard SWBD-DAMSL Shallow Discourse-Function Annotation Coders Manual”, Institute of Cognitive Science Technical Report, pp.1–61 (1997).
(7) Boersma, P.: “Praat, a System for doing Phonetics by Computer”, Glot International, Vol.5, No.9/10, pp.341-345 (2001).
(8) Witten, I.H., Frank, E. and Hall, M.A.: “Data Mining: Practical Machine Learning Tools and Techniques”, 3rd edition, Morgan Kaufmann (2011).