5T-08

ワードウルフにおける発話分類のためのタグセットの提案

脇田尚紀† 菱山玲子†
早稲田大学創造理工学部 経営システム工学科†

1 はじめに

ワードウルフは，少数派と多数派に分かれ対話の中で少数派を見つけだすソーシャルコミュニケーションゲームである．近年，人工知能の発展に伴い囲碁・将棋や人狼ゲームなど様々なゲームを標準問題として研究が行われているが，「新人狼ゲーム」として知名度を伸ばしつつあるワードウルフの研究は現在行われていない．そこで本研究では，ゲームの各プレイヤ間の対話と戦略・勝敗との関係性を解明することで，ワードウルフのプレイヤエージェントモデルを獲得することを最終的な目標とし，対話型ゲームにおける発話分析に有効なタグセットを設計するための新たな手法を提案するとともに，それによって設計されたタグセットの分類精度の評価を行った．

2 タグセット群の獲得プロセス

ワードウルフと同様に少数派と多数派に分かれて対話コミュニケーションによって対戦する人狼ゲームはいくつかの研究が行われている．稲葉ら[1]は以下の 3 つの基準を用いてタグの定義を行った．

1. ゲームの特徴を捉えるタグ
   ゲームの戦略上不可欠な発話に対するタグを定義する．
2. 勝敗への影響の大小による細分化
   意味が類似しているタグであっても勝敗への影響が異なると考えられるタグを分割する．
3. 出現頻度
   勝敗への影響は大きくないと考えられるが，ゲーム中に高頻度で出現する発話に対するタグを定義する．

研究では上記の基準を元に 20 個のタグを定義しているが，タグ設計の性質上個人で設計を行ってしまうと設計されるタグに大きな個人差が生まれてしまうことが考えられる．本研究では個人差による設計のばらつきを抑え，より分類精度の高いタグを設計するために複数人による合議によって定義する方法を提案する．

A Proposal of Tag Sets for Classification of Utterances in Word Wolf Game
Naoki WAKITA†, Reiko HISHIYAMA†
†School of Creative Science and Engineering, Waseda University

3 タグセット獲得実験

本研究ではまずタグを設計する際に参考にするゲームデータを 10 ゲーム分オンライン上で収集する．この時，後の分析を行う際に極力ばらつきが発生しないようにプレイヤの人数は 1 ゲームあたり 4 人，ゲーム時間は 1 ゲームあたり 5 分で統一した．また，ゲーム時間の統一を図るため，サドンデスは行わず，投票多数のプレイヤが複数存在し，かつその中に少数派プレイヤが含まれる場合は引き分けとした．

次に収集した 10 ゲームのデータと稲葉ら[1]のタグ付けの３つの基準を参考に合計 7 人のワードウルフ経験者が発話分類タグセットの提案を行った．それぞれに自由に提案をしてもらうためにタグセットの数等の制限は設けなかった．

最後に集めた発話分類タグセットを元にタグセットを提案した 7 人中 5 人で 1 時間程度の議論を行い，分析に用いるためのタグセット群を決定した．

この実験により表 1 のような 13 個のタグが得られた．

表 1 設計されたタグの一覧

設計されたタグ	タグの定義
Y/N質問(1対1)	「はい」か「いいえ」で答えられる質問(Yes/No Question)を特定の誰かにする
Y/N質問(1対多)	「はい」か「いいえ」で答えられる質問を全員あるいは複数の人にする
O質問(1対1)	「はい」か「いいえ」で答えられない質問(Open Question)を特定の誰かにする
O質問(1対多)	「はい」か「いいえ」で答えられない質問を全員あるいは複数の人にする
答え	質問に対する回答
事実	お題あるいは関連ワードに関する一般的であるとする主張、様子あるいは過去にあった出来事
意見	お題あるいは関連ワードに関する主観による事柄
同意	他の人の発言に対する同意
否定	他の人の発言に対する否定
嘘	自分の持っているお題に則さない発言をする
少数派推測	少数派(ウルフ)のお題や少数派は誰なのかという予測、あるいは特定の誰かに対する疑念
関連ワード	お題に関連する言葉
その他	お題や関連ワードにあまり関係しない意見など、以上の他のタグに分類できないタグ

特定の誰かにされた質問は，相手に対する疑念を確かめるために行われる場合が多々あり，また「はい」か「いいえ」で答えられる質問と答えに幅のある質問では答えから得られる情報量が大きく異なるため，「質問」のタグが 4 つに分けられた．「事実」と「意見」は性質の似たタグであるが，一般的な主張や事象と個人の好き嫌いなどの意見では，プレイヤがその発言に対して自分のお題に則した発言であるかを判断する材料として使いやすさが異なると考えられるため別のタグとして定義された．

4 タグセットの評価

実験で設計されたタグの分類精度を評価するために 2 人のタグ付与者 A,B(以下アノテータ A,B とする)が獲得実験で使用した 10 ゲームに 5 ゲームを加えた 15 ゲーム分のタグ付けをそれぞれ行い，その一致率を確認した．一致率は徳久ら[2]の雑談に対してタグ付けを行った研究の中で提案された下記の式を用いる．

$$\text{一致率}(\%) = \frac{(\text{一致したタグの数})\times 2}{\text{AとBが付与したタグの数}} \times 100$$

表 2 タグの一致率

アノテータAによる付与タグ数	515
アノテータBによる付与タグ数	514
一致タグ数	431
一致率	83.80%

徳久ら[2]の研究で設計されたタグは発話の行為(Dialogue Act)で分類されるタグと修辞構造(Rhetorical Relation)によって分類されるタグの二種類存在するが，その一致率はそれぞれ 65.5%と 59.6%であった．また稲葉ら[1]の研究における一致率は 64.0%であった．このことから本研究のタグセットはより高い一致率が得られていることが分かる．

さらに，cohen の κ を用いた評価も行った．cohen の κ は主観が入る評価に関して二つの評価結果がどの程度一致しているかを調べるための統計量であるが，Landis ら[3]の文献から 0.4〜0.6 の値で中程度の一致，0.6〜0.8 の値でかなりの一致，0.8〜1.0 の値でほぼ一致とされている．cohen の κ は本研究のタグセットのように一つの判定対象に複数の判定結果を与えるような時には適用できないが，今回は稲葉ら[1]の研究と同様に各発話に対して各タグが付与されているかいないかによって判定を行い，cohen の κ の算出を行った．この結果「嘘」以外のタグに関してはかなりの一致とされる 0.6 以上の値が得られたことが分かった．特に「事実」，「同意」，「少数派推測」，「関連ワード」の４つのタグに関してはほぼ一致とされる 0.8 以上の cohen の κ の値を得ることができた．「嘘」のタグに関しても中程度の一致とされる 0.50 の値が得られた．

これを稲葉ら[1]の研究と比較すると 0.8 以上のタグが 20 個中 2 個であった従来研究に対し，本研究で得られたタグは 13 個中 4 個が 0.8 以上の値をとった．また 20 個中 2 個のタグが中程度の一致とされる 0.4〜0.6 の値をとった従来研究に対し，本研究で得られたタグは 13 個中 1 個のタグのみ中程度の一致となった．よって cohen の κ の値からも，本研究の手法によってより分類精度の高いタグセットが得られたことが分かる．

5 考察

今回の評価実験において「嘘」のタグの cohen の κ が低くなってしまった原因としてワードウルフではどちらともとれる表現が多いことが考えられる．例えば「英語(対抗のお題は数学)」というお題を与えられたプレイヤが「ロジックで考えるものなので，感覚派だときついきがします」と発話したのに対しアノテータ A が「嘘」のタグを付与したが，ゲーム後の対話からプレイヤは自分を多数派陣営だと思っており，自分のお題に則さない発話のつもりではないことが明らかであった．よって「嘘」のタグに関しては正確性を担保するため，「嘘」のタグの付与を「社会通念上明らかに与えられたお題に沿わない発言」及び「ゲーム後の対話から発言が対抗への同調のために行われていたと認められるとき」に限定することが望ましいと考える．

6 おわりに

本研究では，オンラインゲーム上で集めたワードウルフのゲームデータを参考に多人数による発話分類タグセットの設計を行った．また，発話タグを設計した他の従来研究と比較することによって，本研究の手法を用いることで，アノテータによる個人差が少なくより分類精度の高いタグが得られることが分かった．今後の課題としては本研究で得られたタグセットを用いてプレイヤの発言傾向や戦略の分析を行うことが挙げられる．

参考文献

[1] 稲葉通将，大畠菜央実，高橋健一，鳥海不二夫：雑談ばかりしていると殺される?人狼ゲームにおける発話行為タグセットの提案とプレイヤの行動・勝敗の分析，情報処理学会論文誌，Vol.57，No.1，pp.2392-2402，2016.

[2] 徳久良子，寺嶌立太：雑談における発話のやりとりと盛り上がりの関連，人工知能学会論文誌，Vol21，No2，pp.133-142，2006.

[3] Landis, J. R., Koch, G. G. :The measurement of observer agreement for categorical data, *biometrics*, pp.159-174, 1977.