雑談対話データへの複数人でのタグ付与における曖昧タグの評価値の推定

杉山 貴昭　　武田 龍　　駒谷 和範
大阪大学 産業科学研究所
{sugiyama@ei., takeda@, komatani@}sanken.osaka-u.ac.jp

1 はじめに

Project Next NLP[1] の対話タスクにおいて，雑談対話データを用いた対話の破綻分析が進められている [7]．このプロジェクトでは，図 1 のようなユーザと雑談対話システム [4] のチャット形式の 10 回のやり取りが 1146 個収集された．アノテータは，システム発話のみ（図 1 の S1 から S10）に対し，○（破綻ではない）・△（破綻と言い切れないが，違和感を感じる）・×（あきらかにおかしい）の 3 種類でタグ付けした．本研究ではこのうち，ランダムに抽出された 100 対話 [7] に対し，24 名のアノテータがタグ付けしたデータを利用する．アノテーションの総数は 24000 個である．

対話が破綻しているか否かの判断は連続値で得られることが望ましい．この理由は，「ほぼ破綻」，「どちらかと言えば破綻ではない」など，破綻には “程度” があるためである．一方で，連続値でのタグ付与は，アノテータへの負担や詳細な基準が必要であるため難しい．そのため，このタスクではシンボル化されたタグ（○△×）が採用された．

この方法では，各タグの解釈に個人差が生じる．図 2 にアノテータ毎の○△×の付与割合を示す．横軸はアノテータの ID であり，左端に全体の○△×の割合（*all*）を示した．図 2 をみると，*all* はおおよそ ○ : △ : × = 6 : 2 : 2 の割合であるのに対し，アノテータによって付与割合は異なることがわかる．つまり，アノテータによって「破綻」の感じ方は異なる．

特に，中間的な評価タグ△はこの影響を最も受ける．このデータ収集では，○△×の付与に具体的な基準は与えられていないため，付与されるタグはアノテータに依存する．そのため，△はそのアノテータにとっての中間的な評価尺度にすぎない．中森らも，採点法で得られる量は，本質的には大小関係のみ意味がある順序尺度 [3] であると指摘している [5]．従って，△はアノテータによって，「どちらかといえば破綻ではない」にも「どちらかといえば破綻」にもなりうる．例えば，図 1 のアノテータ 3 は×を多く付与する傾向があるため，△は「破綻」に近いと解釈できる．また，同様にアノテータ 18 の△は「破綻してない」に近いと解釈できる．このように中間的な評価タグ（以降，曖昧タグ）は人によって解釈が異なる．

本研究では，協調フィルタリングにより，曖昧タグの評価値を推定する方法を提案する．協調フィルタリングを適用すれば，評価値が破綻の程度を表した連続値で得られる．一般的に協調フィルタリングは推薦システムの手法に利用される [1, 2] が，タグの分析において利用された研究はこれまでにない．本モデルが妥当ならば，対話の破綻を連続的な値で表せるため，破綻の度合を考慮したより詳細な分析が可能になる．さらに，アノテータが実際に付与した評価値と推定後に得られた評価値が大きく異なる場合，前者が誤りの可能性があることを，アノテータに指摘する等の利用法も考えられる．

2 協調フィルタリングを利用した曖昧タグの再評価

2.1 協調フィルタリングの利用

本研究では，アノテータによって感じ方が異なる曖昧タグ△を，協調フィルタリングを用いて再評価する．協調フィルタリング [2] とは，ユーザ A と嗜好パターンが類似している他のユーザ B をデータベースから見つけ，それを元にユーザ A にアイテムを推薦する手法である [6]．これは，「ユーザ B と嗜好パターンが類似しているユーザ A は，他のアイテムに対してもユーザ B に似た嗜好パターンを持つ」ことを仮定している．

我々は，協調フィルタリングをアノテータの付与結

[1] https://sites.google.com/site/projectnextnlp/

		アノテータID							
		1	2	3	4	…	18	…	24
S0:	こんにちは！海へ行きたいね。								
U1:	そうだね。								
S1:	いいなー	△	○	△	○	…	○	…	○
U2:	もう海には行った？								
S2:	海は大丈夫です	○	△	×	△	…	○	…	○
⋮	⋮					⋮			
S10:	そんなことないですよ	△	△	×	×	…	△	…	○

図 1: 雑談対話データの例（1407292346.tsv）

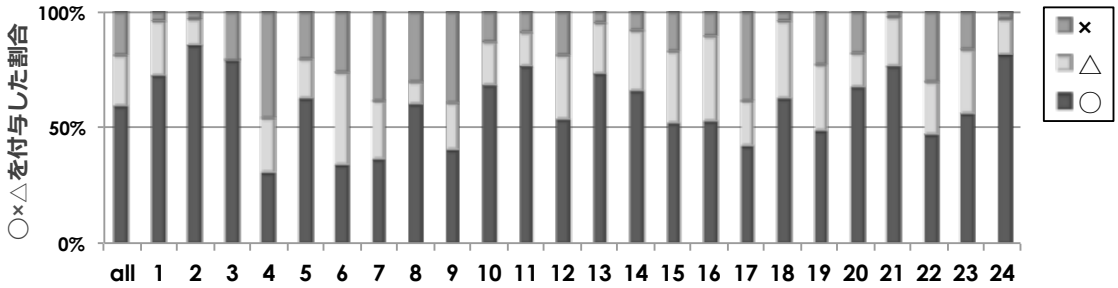

図 2: アノテータ毎の○△×を付与した割合．横軸はアノテータ ID

果の推定に利用する．1 章で述べた通り，曖昧タグ△はアノテータによってその内容の解釈が異なるため，異なるアノテータが付与した△を同様に扱うことは問題である．そこで，協調フィルタリングを利用し，△の評価値の推定を試みる．本研究では,「アノテータ A と付与の傾向が類似したアノテータ B は，他のデータに対してもアノテータ A に似た傾向で付与する」ことを仮定する．これを仮定すれば，例えば，図 1 のアノテータ 2 と 4 の付与結果が類似していることから，アノテータ 2 の S10 に対する付与結果△は「破綻してない」に近いと考えられる．

2.2 曖昧タグの評価値の推定

行列分解に基づく手法 [1] を用いて，曖昧タグの評価値を推定する．アノテータ $n(n = 1, ..., 24)$ の発話 $m(m = 1, ..., 1000)$ の評価値を r_{nm} とする．推定後の評価値 $\hat{r}_{nm}$ は潜在ベクトルの内積の近似で表せる．

$$\hat{r}_{nm} \approx \mathbf{u}_n^{\mathrm{T}} \mathbf{v}_m = \sum_{k=1}^{K} u_{nk} v_{mk} \tag{1}$$

$\mathbf{u}_n^{\mathrm{T}}$ はアノテータ n の潜在ベクトルを，$\mathbf{v}_m$ は発話 m の潜在ベクトルを表す．K は次元数であり，潜在ベクトルがアノテータや発話からどの程度特徴を捉えるかを表す．本稿では，まずアノテータ毎の特性を表現できる $K = 24$ とした．

次に，評価値 $\hat{r}_{nm}$ を用いて，目的関数 E を計算する．

$$\begin{aligned} E = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} I_{nm} ||r_{nm} - \hat{r}_{nm}||^2 \\ + \frac{\lambda}{2} \sum_{n=1}^{N} ||\mathbf{u}_n||^2 + \frac{\lambda}{2} \sum_{m=1}^{M} ||\mathbf{v}_m||^2 \end{aligned} \tag{2}$$

I_{nm} は評価値の有無を表し，○か×の時は 1，△の時は 0 である．つまり，アノテータ n の発話 m に対する付与結果が△の箇所を他のタグから推定している．r_{nm} はアノテータが実際に付与した評価値であり，○は 1，△は 0，×は −1 とした．λ は正則化係数である．

この目的関数を最小化する $\mathbf{u}$ と $\mathbf{v}$ を推定する．更新式は下記の通りである．

$$\mathbf{u}_n \leftarrow \mathbf{u}_n - \eta \frac{\partial E}{\partial \mathbf{u}_n} \tag{3}$$

$$\mathbf{v}_m \leftarrow \mathbf{v}_m - \eta \frac{\partial E}{\partial \mathbf{v}_m} \tag{4}$$

η は，学習係数である．終了条件は，$\Delta E < 0.1$ とした．また，$\mathbf{u}_n^{\mathrm{T}}$ と $\mathbf{v}_m$ の初期ベクトルは，−1 から 1 の乱数とした[2]．

[2]初期ベクトルによって最終的に得られる推定値が異なるため，最適な初期ベクトルの設定は今後の課題である．

発話ID m	アノテータID n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
4		○	△	○	○	△	○	△	×	△	○	○	△	×	○	○	△	○	○	○	△	○	△	△	○
		0.76	0.41	0.96	0.99	1.16	1.06	0.26	−0.96	−0.06	0.91	0.92	0.10	0.24	0.89	1.00	0.87	0.92	1.06	1.02	0.83	0.56	0.44	1.10	0.60
30		○	○	○	○	○	○	○	○	○	○	○	○	○	○	△	○	○	○	○	○	○	○	○	○
		1.01	0.97	1.06	0.98	1.12	0.87	1.01	1.04	0.97	1.08	0.94	1.01	0.92	1.12	0.93	0.92	0.97	1.00	0.99	1.08	0.99	1.03	0.86	0.97
67		△	○	×	×	×	△	×	△	×	△	○	○	○	△	×	△	×	△	△	×	○	△	△	○
		1.09	1.17	−0.81	−0.97	−0.89	−0.88	−0.99	0.10	−0.97	−1.10	0.77	0.94	0.83	−0.15	−1.05	0.29	−0.89	0.68	−0.85	−0.88	1.02	0.04	−0.80	0.88
76		×	△	×	×	×	×	×	×	×	×	×	△	×	△	×	×	×	△	△	×	×	×	×	×
		−1.03	−0.91	−1.00	−0.94	−1.07	−0.93	−1.01	−1.00	−0.97	−1.02	−0.96	−0.86	−0.94	−1.16	−0.96	−0.92	−1.00	−0.98	−1.71	−1.05	−0.99	−0.99	−0.98	−0.94
269		△	○	○	△	×	△	△	×	×	×	○	△	△	○	○	△	×	△	○	△	△	×	×	○
		0.95	1.12	0.87	−1.09	−1.00	−0.35	0.79	−0.99	−0.96	−0.97	0.87	0.23	1.23	0.95	0.72	0.91	−0.82	1.07	0.71	0.86	1.10	−1.05	−0.55	0.84
375		△	△	○	×	×	×	△	×	×	△	×	△	○	○	△	△	×	○	△	○	○	×	×	○
		0.80	1.16	0.82	−1.06	−1.19	−0.86	−1.19	−1.01	−0.90	0.16	−0.69	−0.20	0.99	0.95	−0.25	−0.20	−0.92	0.87	−0.75	0.78	0.93	−1.11	−0.76	0.95

図 4: アノテータの付与結果と本モデルで推定した評価値の例

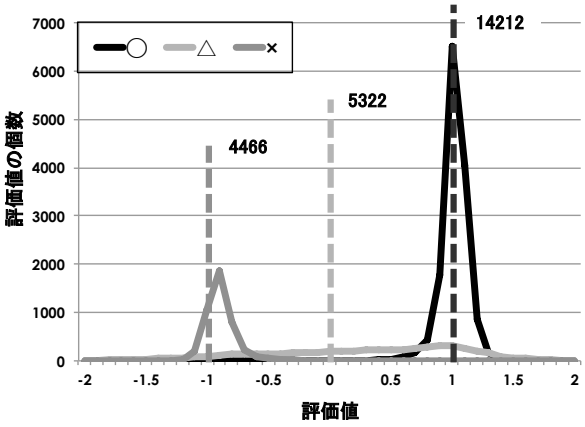

図 3: 推定後の評価値の分布

3 モデルの妥当性と利用方法に関する調査

3.1 モデルの妥当性の検証

協調フィルタリングによる推定後の評価値の分布を，図 3 のヒストグラムに示す．横軸が評価値であり，縦軸が評価値の個数である．なお，評価値は 0.1 毎に集計した．また，実線が推定後の分布，点線が推定前の分布である．図 3 を見ると，○と×は推定後も 1，−1 付近に多く分布しているのに対し，△は散らばっていることがわかる．実際に，推定後の評価値の最大値は○が 1.5，△が 2.4，×が 0.7 であり，最小値は○が −0.4，△が −2.5，×が −1.4 だった．

曖昧タグの推定後の評価値が，アノテータや発話の特性に基づき，適切な値になっていることを調べる．図 4 に，アノテータの付与結果のうち 6 発話分[3]を例として取り出して示す．アノテータの付与結果の下には，本モデルにより推定した評価値 $\hat{r}_{nm}$ も記載した．

まず，発話の潜在ベクトルについて調べる．$m = 30$ の発話に対する付与結果をみると，ほとんどの人が○を付与し，$n = 15$ のアノテータだけ△を付与したことがわかる．このとき，推定後の評価値 $\hat{r}_{15,30}$ は 0.93 だった．逆に，$m = 76$ の発話に対する $n = 2$ のアノテータの評価値 $\hat{r}_{2,76}$ は −0.91 だった．実際の対話は以下であり，$m = 76$ は下線部である．

S5: 都合が良いですね
U6: 暇になったので遊びに行きたかったのですよ
S6: 都合が政府に悪いですよねぇ
U7: それは少し違うと思います．

人間が見ても，$m = 76$ は「政府」が唐突に現れており，対話が破綻していると判断する可能性が高い．つまり，推定後の評価値は，他のアノテータが付与したタグの比によって変化している．また，図 4 の $m = 375$ の発話の場合，アノテータ $n = 7$ の△の評価値 $\hat{r}_{7,375}$ は −1.19 だった．これは，発話の潜在ベクトルが，付与の傾向が類似しているアノテータの付与結果も考慮しているためだと考えられる．例えば，図 2 を見ると，アノテータ $n = 7$ と $n = 4$ の付与割合は類似している．そのため，$\hat{r}_{7,375}$ の評価値の推定に，$n = 4$ の付与結果が考慮されている可能性が高い．従って，発話の潜在ベクトルは，発話毎のタグの比や類似したアノテータの付与結果を学習していることが示されている．

次に，アノテータの潜在ベクトルが，アノテータ毎の○△×を付与した割合を考慮していることを確かめる．図 4 の $m = 67$ の発話では，$n = 10$ のアノテータの付与結果の評価値 $\hat{r}_{10,67}$ は −1.10 だった．実際の対話は以下であり，$m = 67$ は下線部である．

S6: 人口は多いです
U7: 多いですね。待ちは混んでますしね。
S7: 待ちは良いですね
U8: 良いですかね？

この文脈を見ると，U7 のタイプミスをシステムがそ

[3]スペースの関係でこの 6 つのみを掲載した．

のまま出力したため，アノテータの評価が割れている．従って，この発話に対する評価は，アノテータの特性に依存すると考えられる．図 2 を見ると，このアノテータは◯を付与する割合が多い．よって，このアノテータの△は「破綻してない」に近いと解釈できるため，△の評価値が -1 に近くなったと考えられる．逆に，$m = 269$ の発話に対する $n = 7$ のアノテータの評価値 $\hat{r}_{7,269}$ は 0.79 だった．これは，このアノテータは×を多く付与する傾向があるため，△の評価値が 1 に近くなったと考えられる．以上より，アノテータの潜在ベクトルは，アノテータ毎の◯△×を付与した割合を学習していることを示した．

3.2 推定結果の利用方法の検討

アノテータの付与誤りの訂正

本モデルの推定結果がアノテータの付与誤りの訂正に利用できることを示す．まず，3.1 節で述べたように，推定後の◯の最小値が -0.4 であり，×の最大値が 0.7 であることから，評価が逆転したデータが存在していることがわかる．このようなデータは，アノテータが付与を誤っている可能性が高い．

我々は，推定前後の評価値の差 $d_{nm} = |r_{nm} - \hat{r}_{nm}|$ からアノテータの付与誤りを検出できると考える．例えば，図 4 の $m = 4$ の発話の場合，◯の付与が多いアノテータ $n = 13$ の付与結果は×で，その評価値 $\hat{r}_{13,4}$ は 0.24 だった．図 2 よりこのアノテータは◯を多く付与するアノテータであり，他のアノテータも◯を付与しているにも関わらず，×を付与していた．このような場合，$d_{13,4} = |-1.0 - 0.24| = 1.24$ となり，推定前後の差が大きくなる．実際の $m = 4$ の発話は下記の通りであり，$m = 4$ を下線で示す．

S3: クーラーは快適ですね
U4: うん、会社だと朝はつけられないから死にそう
S4: 会社は冷房が効きますよねぇ
U5: 昼間は効くね。

実際の対話を見ても，S4 のシステム発話によって，対話が破綻したとは言い難い．実際にユーザはうまく返答（U5）できている．このような付与結果は，アノテータの付与誤りである可能性が高い．そこで，d_{nm} が大きい付与結果をアノテータに指摘すれば，このようなエラーを防ぐことができる．

対話の破綻の程度を考慮した分析

1 章で述べた通り，アノテーション時にシンボル化されたタグを用いることで，アノテータの負担を軽減できる．一方で，対話の破綻には “程度” があるため，連続値で分析できることが望ましい．本モデルで推定した評価値は，破綻の程度を表した連続値であるため，実際の対話の破綻分析に利用可能である．もし△箇所の評価値が 0.9 であれば,「破綻していない」，-0.4 であれば,「どちらかと言えば破綻」といったような解釈が可能になる．

4 おわりに

本稿では，協調フィルタリングに基づき曖昧タグの評価値を推定する手法を提案した．本モデルにより得られた評価値が，破綻の程度を表した連続的な値で得られることを示した．また，本モデルが実際の対話分析や，アノテータの付与誤りの訂正に利用できることを述べた．

今後の展開として，最適な潜在ベクトルの次元数の設定が挙げられる．本稿では $K = 24$ と設定したが，これはアノテータの人数と同数であるため，各アノテータに対し過度に適応している可能性がある．そこで，特異値分解で潜在ベクトルの累積寄与率を算出するなどして，最適な潜在ベクトルの次元数を決定する方法を検討している．

参考文献

[1] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, Vol. 42, No. 8, pp. 30–37, 2009.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 175–186, 1994.

[3] S. S. Stevens. *Mathematics, measurement, and psychophysics*. 1951.

[4] 大西可奈子, 吉村健. コンピュータとの自然な会話を実現する雑談対話技術. NTT DoCoMo テクニカル・ジャーナル, Vol. 21, No. 4, pp. 17–21, 2014.

[5] 中森義輝. 感性データ解析ー感性情報処理の為のファジィ数量分析手法. 森北出版, 2000.

[6] 神嶌敏弘. 推薦システムのアルゴリズム (2). 人工知能学会論文誌, Vol. 23, No. 1, pp. 89–103, 2007.

[7] 東中竜一郎, 船越孝太郎. Project Next NLP 対話タスクにおける雑談対話データの収集と対話破綻アノテーション. 人工知能学会研究会資料, Vol. SIG-SLUD-B402, pp. 45–50, 2014.