チャットシステムのための共感発話の推定

福岡 知隆　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1320010, kshirai}@jaist.ac.jp

1 はじめに

対話システムには大きく分けて 2 つの種類がある．1 つはタスク指向型対話システムである．道案内，ホテルの予約，商品のプロモーションなど，特定のタスクを設定し，対話を行う．もう 1 つは自由対話システムである．これはタスクを限定せず，様々な話題について自由に対話を行うシステムである．従来の研究はタスク指向型対話システムが中心であったが，近年はいやしを求めるペット型ロボットや介護ロボットなど，自由対話システムの重要性が増している．

自由対話における重要な要素の 1 つに，対話の内容に対する話者の共感が挙げられる．自由対話では話題は固定されておらず，任意のタイミングで変更することが可能である．しかし，話題を変更するタイミングはいつでも良いわけではない．相手がまだその話題について話をしたい時に話題を変更したり，またはその話題についてはこれ以上対話を続けたくないにも関わらず対話を続けることは，相手に不快感を与え，場合によっては対話を打ち切られる可能性がある．この話題転換のタイミングを計る要素の 1 つが話者の共感である．話し相手の共感が得られているならば現在の話題を継続し，逆に共感が得られていないならば異なる話題を提供することが，自然な自由対話を成立させるために重要である．

本論文は，自由対話を行うチャットシステムにおける要素技術として，ユーザの発話が共感を示しているか否かを推定する手法を提案する．人間が対話中に相手の共感を得られているかを判断する手がかりとしては，相手の表情，しぐさ，発話の内容などが考えられる．我々は，この中で発話内容の処理に重点を置き，書き起こされた発話のテキストから共感の有無を判断する．

2 関連研究

一般に，対話の自動タグ付けには機械学習がよく使われる [1]．一方，共感の有無も対話におけるタグの一種と考えられる．本研究でも機械学習を用いて共感の有無を判断する．

Bo らは単語の n-gram を素性とし，言語モデル学習ツール SRILM を用いて共感発話の推定を行った [2]．彼らは，bi-gram を素性とする結果が最も精度が高く，正答率 6 割程度であることを示した．南らは，共感を含む 29 の対話行為タイプを定義し，その自動認識を発話中の単語を利用する重み付け有限状態トランスデューサーを用いて行った [3]．この実験では対話行為タイプの認識精度は 50 %であった．関野らは，条件付き確率場 (Conditional Random Field; CRF) を用いて対話行為タグの自動付与を行った．対話行為タグの定義は SWBD-DAMSL タグセット [5] を用いており，これは sympathy(共感) というタグを含む．CRF の学習素性として，直前の対話タグ，内容語数，発話長を用いている．

本論文では，これらの関連研究を踏まえて，ユーザの発話が共感を示すかを判断するための機械学習の素性を考案する．また，学習素性の有効性を実験により評価する．

3 提案手法

本節では，自由対話のテキストを入力とし，発話毎にその発話が共感を示しているかを推定する手法を提案する．共感発話の推定のために教師あり機械学習を用いる．具体的には，メモリーベース学習手法の 1 つである Timbl[6] と Support Vector Machine(SVM) の 2 つの手法で共感の有無を判定する二値分類器を学習する．また，4 節では両者を実験で比較する．なお，本論文における共感発話とは，相手の発話を受けて共感・賛意を示す発話とする．

以下，提案手法における学習素性について述べる．

f_1：**発話長** 共感を示す発話は文長が短い傾向が見られるため，発話長を学習素性として用いる．発話長を Timbl や SVM の素性として用いる場合，長さを適当な間隔（1 ～ 5，6 ～ 10，11 以上，など）に切って発話長を分類するのが一般的であるが，その適切な間隔を決めるのは難しい．本論文では，「発話長が $l \pm 2$ である」$(3 \leq l \leq 20)$,「発話長が

23 以上である」という 19 種類の素性で発話長を表現する．

f_2：**現在と直前の発話の文末 n-gram** 「〜だね」「〜よね」という文末は共感を示唆するように，共感を示す発話は文末に特徴がある．また，発話者 A が「〜だよね」と共感を求める発話に対して発話者 B が「そう思う」と応じるように，直前の発話者の文末にも特徴がある．ここでは，現在の発話の文末の単語 n-gram と直前の発話の文末の単語 n-gram(ただし，$n = 1, 2, 3$) の組み合わせを素性とする．なお，予備実験では，現在ならびに直前の発話の文末 n-gram を別々の素性としたときも試したが，結果は良くなかった．

f_3：**発話者変更** 発話者の変更が共感を示す発話の出現に影響があると考えられるため，発話者に変更があったか否かを素性とする．

f_4：**自立語繰り返し 1** 発話者 A が「あれは傑作だった」と言ったのに対して話者 B が「傑作だね」と応じるように，相手の発話中の語を繰り返して共感を示すときがある．そのため相手の直近の発話中の自立語を含むか否かを素性とする．

f_5：**自立語繰り返し 2** f_4 と同じ考えに基づくが，共感を示す自立語の繰り返しをより厳密に定義する．具体的には,「相手の直近の発話の最後の用言が発話中に存在するか」と,「発話に出現する自立語が 1 つのみであり，かつそれが相手の直近の発話に出現するか」という 2 種類の素性を用いる．

f_6：**共感を表わす発話** 「そう。」,「あるある。」のように短い一文だけで共感を表わす場合がある．ここでは共感を示す 21 の発話を用意し，発話がそのいずれかに該当するかを素性とする．なお，共感を表わす 21 の発話は 4.1 項で述べる実験データを参照して人手で作成した．

f_7：**他の対話タグ** 対話行為は共感の有無を判定する有力な手がかりである．例えば，共感は相手の主張や考えなどに対して示されることが多いが，質問に対して共感することは少ない．本論文では，自己開示，質問（YesNo），質問（What），応答（YesNo），応答（平叙），あいづち，フィラー，確認，要求の 9 種類の対話行為を定義し，これを素性とした．同様に，発話の主観性を表わすタグとして，主観的，客観的，どちらでもない，という 3 種類を定義し，これも素性とした．なお，実験ではこれらの対話タグは人手で与えた．対話タグを自動分類し，これを用いて共感の有無を判定するモデルを学習することは今後の課題である．

4 評価実験

本節では提案手法の評価実験について述べる．

4.1 使用コーパス

実験には名大対話コーパス [7] を用いた．このコーパスは実際の雑談を書き起こしたものである．名大対話コーパスから対話参加人数が二人の対話を 4 つ選び，対話内の個々の発話に対して共感を示しているかのタグを人手で付与した．表 1 に実験に用いた対話の発話数ならびに共感ありのタグが付与された発話数を示す．

表 1: 実験データ

対話番号	発話数	共感発話数
003	718	57
004	879	31
007	797	15
009	1618	33

4.2 実験方法

1 つの対話をテストデータ，それ以外を訓練データとする交差検定法により，共感の有無を判定する．分類器を学習するメモリーベース学習には Timbl を用い，SVM の学習には libSVM[8] を用いた．学習素性の有効性を測るため，全ての素性を利用した場合 (all) と，学習素性を 1 つだけ除いた場合 ($all - f_i$) について，共感発話の判定の精度，再現率，F 値を求めた．

4.3 実験結果

メモリーベース学習と SVM のそれぞれについて，共感判定の精度を表 2 に，再現率を表 3 に，F 値を表 4 に示す．各表には 4 つの対話の結果およびその平均を示した．表 2，3 における括弧内の数値は，精度ならびに再現率の計算式における分子と分母である．

4.4 考察

精度と再現率を比較すると，全体的に精度の方が高い傾向が見られた．全ての素性を使ったとき，メモリーベース学習の精度，再現率はそれぞれ 0.37，0.26，SVM の精度，再現率は 0.81，0.21 であった．メモリーベース学習と SVM を比較すると，精度は SVM の方が高く，再現率はメモリーベース学習の方が高い．ただし，全ての素性を使ったときの F 値を比較すると，

表 2: 共感判定の精度

	003	004	007	009	平均
メモリーベース学習					
all	0.63 (17/27)	0.26 (5/19)	0.21 (4/19)	0.31 (10/32)	0.37
$all - f_1$	0.73 (16/22)	0.31 (5/16)	0.27 (4/15)	0.23 (9/39)	0.37
$all - f_2$	0.64 (16/25)	0.26 (5/19)	0.25 (4/16)	0.37 (10/27)	0.40
$all - f_3$	0.65 (17/26)	0.31 (5/16)	0.17 (4/24)	0.38 (11/29)	0.39
$all - f_4$	0.64 (16/25)	0.31 (5/16)	0.22 (4/18)	0.36 (10/28)	0.40
$all - f_5$	0.67 (16/24)	0.35 (9/26)	0.24 (5/21)	0.34 (11/32)	0.40
$all - f_6$	0.43 (10/23)	0.28 (7/25)	0.00 (0/7)	0.29 (12/42)	0.30
$all - f_7$	0.57 (8/14)	0.25 (3/12)	0.28 (8/29)	0.09 (3/32)	0.25
SVM					
all	0.54 (7/13)	0.93 (14/15)	1.00 (6/6)	1.00 (2/2)	0.81
$all - f_1$	0.67 (8/12)	0.92 (11/12)	0.86 (6/7)	0.67 (2/3)	0.79
$all - f_2$	0.69 (9/13)	0.81 (17/21)	0.67 (6/9)	0.50 (5/10)	0.70
$all - f_3$	0.55 (6/11)	0.93 (13/14)	1.00 (6/6)	1.00 (2/2)	0.82
$all - f_4$	0.57 (8/14)	0.93 (14/15)	1.00 (6/6)	1.00 (2/2)	0.81
$all - f_5$	0.50 (6/12)	0.94 (15/16)	1.00 (6/6)	0.67 (2/3)	0.78
$all - f_6$	0.50 (8/16)	0.92 (11/12)	1.00 (6/6)	0 (0/0)	0.74
$all - f_7$	0.50 (1/2)	0 (0/0)	1.00 (2/2)	0.28 (5/18)	0.36

表 3: 共感判定の再現率

	003	004	007	009	平均
メモリーベース学習					
all	0.30 (17/57)	0.16 (5/31)	0.27 (4/15)	0.30 (10/33)	0.26
$all - f_1$	0.28 (16/57)	0.16 (5/31)	0.27 (4/15)	0.27 (9/33)	0.25
$all - f_2$	0.28 (16/57)	0.16 (5/31)	0.27 (4/15)	0.30 (10/33)	0.26
$all - f_3$	0.30 (17/57)	0.16 (5/31)	0.27 (4/15)	0.33 (11/33)	0.27
$all - f_4$	0.28 (16/57)	0.16 (5/31)	0.27 (4/15)	0.30 (10/33)	0.26
$all - f_5$	0.28 (16/57)	0.29 (9/31)	0.33 (5/15)	0.33 (11/33)	0.30
$all - f_6$	0.18 (10/57)	0.23 (7/31)	0.00 (0/15)	0.36 (12/33)	0.21
$all - f_7$	0.14 (8/57)	0.10 (3/31)	0.53 (8/15)	0.09 (3/33)	0.16
SVM					
all	0.21 (7/33)	0.25 (14/57)	0.19 (6/31)	0.13 (2/15)	0.21
$all - f_1$	0.24 (8/33)	0.19 (11/57)	0.19 (6/31)	0.13 (2/15)	0.20
$all - f_2$	0.27 (9/33)	0.30 (17/57)	0.19 (6/31)	0.33 (5/15)	0.27
$all - f_3$	0.18 (6/33) 0.23	0.23 0.19	0.19	0.13 (2/15)	0.20
$all - f_4$	0.24 (8/33)	0.25 (14/57)	0.19 (6/31)	0.13 (2/15)	0.22
$all - f_5$	0.18 (6/33)	0.26 (15/57)	0.19 (6/31)	0.13 (2/15)	0.21
$all - f_6$	0.24 (8/33)	0.19 (11/57)	0.19 (6/31)	0.00 (0/15)	0.18
$all - f_7$	0.03 (1/33)	0.00 (0/57)	0.06 (2/31)	0.33 (5/15)	0.06

表 4: 共感判定の F 値

	003	004	007	009	平均
メモリーベース学習					
all	0.40	0.20	0.24	0.31	0.31
$all - f_1$	0.41	0.21	0.27	0.25	0.30
$all - f_2$	0.39	0.20	0.26	0.33	0.31
$all - f_3$	0.41	0.21	0.21	0.35	0.32
$all - f_4$	0.39	0.21	0.24	0.33	0.31
$all - f_5$	0.40	0.32	0.28	0.34	0.34
$all - f_6$	0.25	0.25	0.00	0.32	0.25
$all - f_7$	0.23	0.14	0.36	0.09	0.20
SVM					
all	0.30	0.39	0.32	0.24	0.34
$all - f_1$	0.36	0.32	0.32	0.22	0.32
$all - f_2$	0.39	0.44	0.30	0.40	0.39
$all - f_3$	0.27	0.37	0.32	0.24	0.32
$all - f_4$	0.34	0.39	0.32	0.24	0.35
$all - f_5$	0.27	0.41	0.32	0.22	0.34
$all - f_6$	0.33	0.32	0.32	0.00	0.29
$all - f_7$	0.06	0.00	0.12	0.30	0.10

メモリーベース学習は 0.31，SVM は 0.34 であることから，SVM の方が共感発話の判定に適していると言える．

共感発話推定の精度，再現率は十分に高いとは言えない．特に再現率が低い．これは，共感を持つ発話の数（正例の数）が少ないことが大きな要因と考えられる．表 1 より，対話によって異なるが，正例の占める割合は 2〜8%程度で，4 つの対話全体では 3.4%である．

次に，個々の素性の有効性について考察する．素性集合から 1 つの素性を除いた実験結果から，全ての素性を用いたときと比べて評価指標の値が大きく下がるのは f_6 と f_7 で，これらの素性が共感発話の判定に特に有効であることがわかる．逆に，素性を除くと評価指標が向上し，その素性が判定に悪影響を与えると考えられる場合も見られた．特に，メモリーベース学習での f_5，SVM での f_2 は全ての素性を使った場合との差が大きかった．

f_6「共感を表わす発話」に着目すると，all は $all - f_6$ と比べて，共感ありと判定した文の数が増えるが，正答数も増え，結果として精度，再現率ともに向上している．ただし，共感を表わす発話のリストは実験に用いた対話データから取得しているため，この素性についてはクローズドテストであり，有効に働くのは当然と言える．ただし，メモリーベース学習での対話 004 や

009，SVM での対話 003 といったように，対話によっては f_6 を加えても F 値が改善しない (all が $all - f_6$ よりも低い) 場合もある．これは，対話の内容によって共感を表わす発話が異なるためと考えられる．

f_7「他の対話タグ」に着目すると，all は $all - f_7$ を上回り，さらにその差は all と $all - f_6$ の差よりも大きい．したがって，自己開示や質問などの対話行為のタグ，あるいは発話の主観性のタグは f_6 よりも有効な素性であると言える．ただし，メモリーベース学習での対話 007 は，例外的に素性 f_7 を加えても F 値が改善しなかった．

メモリーベース学習における f_5「自立語繰り返し 2」の効果を検証すると，all は $all - f_5$ と比べて正答数が減り，精度，再現率ともに低下している．f_4「自立語繰り返し 1」も大きな効果がなかったことから，自立語の繰り返しという素性はメモリーベース学習では有効に働かなかった．

SVM における f_2「文末の n-gram」の効果を検証すると，all は $all - f_2$ と比べて，誤検出が減ったために精度は向上したが，正答数も大きく減ったために再現率は低下し，F 値も低くなっている．精度は向上することから悪影響を与えるだけの素性とは言えないが，素性集合が $all - f_2$ のときの SVM の F 値は 0.39 で，今回の実験結果の中では最も高かった．

同じ素性でも，対話によって，評価指標の向上に貢献するときと逆に低下を招くときがある．これは，共感発話の特徴，ひいては共感の自動判定に有効な素性が対話の内容によって異なる可能性があることを示唆する．機械学習を用いる場合，対話の内容も考慮し，例えば話題が似ている対話を訓練データとするなどの工夫が必要かも知れない．

5 おわりに

本論文では発話テキストから得られる素性を用いた機械学習に基づく共感発話の推定手法を提案した．評価実験の結果から，共感発話の推定には，共感を表わす発話，および対話行為や主観性といった対話タグを素性とすることが有効であるという知見が得られた．ただし，判定の F 値が全体的に低いこと，対話の内容によって素性の有効性が異なることから，共感発話の推定に有効な素性の検討は十分であるとは言えない．また，素性の組み合わせの有効性についての検討も必要である．

今後の課題として，訓練データにおける正例が少ないという問題を克服する必要がある．例えば，簡単な規則によって明らかに共感を持たない発話を除去することで，正例と不例の数の不均衡を是正する方法などを検討している．

参考文献

[1] 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，”電子情報通信学会論文誌. A, 基礎・境界, vol.92, no. 11, pp. 795-805， 2009.

[2] Bo Xiao, Dogan Can, Panayiotis G. Georgiou, David C. Atkins and Shrikanth Narayanan. “Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy,” Interspeech, 2012.

[3] 南泰浩，東中竜一郎，堂坂浩二，目黒豊美，森啓，前田英作．“対話行為タイプ列 Trigram による行動予測確率に基づく POMDP 対話制御，” 電子情報通信学会論文誌. A, 基礎・境界，vol.95, no.1, pp. 2–15, 2012.

[4] 関野嵩浩，井上雅史，横山晶一，“発話に対する拡張談話タグ付与，” 第 6 回情報処理学会東北支部研究会報告，2010.

[5] D. Jurafsky, E. Shriberg, D. Biasca. “Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual,” Draft 13, University of Colorado, Institute of Cognitive Science, Tech. Rep, pp. 97-101, 1997.

[6] W Daelemans, J Zavrel, K Van der Sloot, and A Van den Bosch. “TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide,” ILK Research Group Technical Report Series no. 10-01. 2010.

[7] 名大会話コーパス，科学研究費基盤研究 (B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成 13 年度〜15 年度）

[8] Chang, Chih-Chung and Lin, Chih-Jen. “LIBSVM: A library for support vector machines,” ACM Transactions on Intelligent Systems and Technology, vol. 65, no. 3, pp. 1–27, 2011.