特集論文 「知的対話システム」

オンライン上の対話における聞き手の感情の予測と喚起

Predicting and Evoking Listener's Emotion in Online Dialogue

長谷川 貴之 Takayuki Hasegawa
東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo
hasegawa@tkl.iis.u-tokyo.ac.jp, http://www.tkl.iis.u-tokyo.ac.jp/~hasegawa/

鍛治 伸裕 Nobuhiro Kaji
東京大学 生産技術研究所
Institute of Industrial Science, the University of Tokyo
kaji@tkl.iis.u-tokyo.ac.jp, http://www.tkl.iis.u-tokyo.ac.jp/~kaji/

吉永 直樹 Naoki Yoshinaga
（同 上）
ynaga@tkl.iis.u-tokyo.ac.jp, http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/

豊田 正史 Masashi Toyoda
（同 上）
toyoda@tkl.iis.u-tokyo.ac.jp, http://www.tkl.iis.u-tokyo.ac.jp/~toyoda/

keywords: emotion analysis, response generation, dialogue

Summary

While there have been many attempts to estimate the emotion of a speaker from her/his utterance, few studies have explored how her/his utterance affects the emotion of the listener. This has motivated us to investigate two novel tasks: predicting the emotion of the listener and generating a response that evokes a specific emotion in the listener's mind. We target Japanese Twitter posts as a source of dialogue data and automatically build training data for learning the predictors and generators. The feasibility of our approaches is assessed by using 1099 utterance-response pairs that are built by five human workers.

1. はじめに

円滑な対話を行うためには，相手の感情に配慮することが重要となる．例えば，落ち込んでいる相手に励ましの言葉をかけることや，相手の感情を害するような発言を避けることの重要性は，改めて説明するまでもないことであろう．

これまで，音声認識や自然言語処理をはじめ人工知能における諸研究分野においては，対話における人の感情に着目した研究が数多く行われてきた [Ayadi 11, Bandyopadhyay 11, Balahur 11]．これらの研究においては，例えば話し手の感情推定に代表されるように，発話内容と話し手の感情の関係性に主眼が置かれてきた．しかしながら，その一方で，発話内容が聞き手の感情にどのような影響を与えるのか，という観点からの研究は十分に行われていない．

本論文では，対話における聞き手の感情に着目したタスクとして「聞き手の感情予測」「聞き手の感情を喚起する発話応答生成」という 2 つのタスクを提案し，それらを実現する方法について述べる．

まず，感情予測タスクでは，システムは与えられた対話履歴に基づいて聞き手の感情を出力する．本研究では，単純化のため，対話の参与者は 2 名とする (図 1)．そして，感情予測の対象となる人物をターゲット，その対話相手をパートナーと呼ぶ．対話履歴としては，ターゲットの発話とそれに対するパートナーの応答が与えられるとする．そして，パートナーからの応答を受け取ったターゲットが抱く感情を予測する．例えば，図 1 上のような対話履歴に対しては「喜び」と予測し，また図 1 下の場合には「悲しみ」と予測する (図 1)．

次に，応答生成タスクでは，システムにはターゲットの発話と「喜び」や「悲しみ」のような感情カテゴリ (目的感情と呼ぶ) が与えられる．そして，ターゲットに目的感情を喚起する応答を生成する*1．例えば，目的感情を「喜び」とした場合,「3 日間高熱が出ているの」という入力に対しては「すぐに良くなるといいね」のような応答を生成する．また，目的感情が「悲しみ」の場合には「風邪が治るまで来ないでね」のような応答を生成する (図 1)．

こうしたタスクを設定して解くことには，以下のような工学的意義が考えられる．ある発話によって聞き手に喚起される感情を計算機で予測することが可能になれば，例えばカスタマサービスセンターなどにおいて，聞き手 (=

*1 このタスクではシステムがパートナーの役割を演じる．

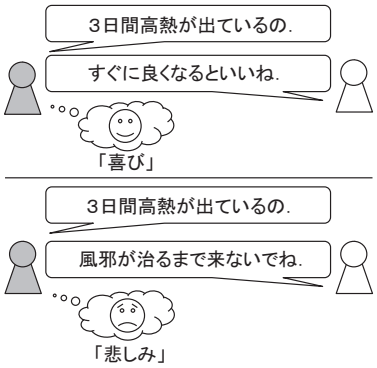


図 1 ターゲット (左側) の発話に対するパートナー (右側) の応答例．これらの応答はそれぞれ「喜び」と「悲しみ」をターゲットに喚起している．

顧客) が気分を害するような発話を自動的に検出し，フィルタリングするようなシステムへ応用が想定できる．同様のアプリケーションは，メールなどのオンライン上でのやりとりにおいても有用であると考えられる [Spertus 97]．また，聞き手の感情を喚起する応答生成が可能になれば，チャットやマイクロブログなどのオンラインコミュニケーション環境における，入力支援技術の高度化につながることが期待できる [Hasselgren 03, Pang 12].

上記のようなタスクを実現しようとした時，最も問題となるのが，辞書や訓練コーパスといった言語資源の構築である．聞き手に感情を喚起させる言語表現は多種多様であり，そうした表現を計算機によって認識可能とするためには，大規模な言語資源が必要になると考えられる．しかしながら，そのような言語資源をどのようにして構築すれば良いのかは自明なことではない．これは，知識獲得ボトルネックとして，人工知能や自然言語処理において広く指摘されている問題である．

この問題に対処するため，本研究ではマイクロブログ上の対話データに着目する．近年，マイクロブログは社会に浸透し，そこでは多数のユーザが対話的なやりとりを行っている．これにより，大規模な対話データをウェブから容易に取得することが可能になりつつある．そこで，少数の手がかり表現を利用することによって，そうした対話データから大規模な感情タグ付きコーパスの自動構築を行う．このようなアプローチを取ることによって，人手でコーパス構築を行うという自明な方法と比べ，小さな作業コストで大規模なコーパスを構築することが可能となる．

感情予測タスクにおいては，上記のコーパスを利用して分類器の学習を行う．感情予測タスクと関連が深いタスクとしては，ニュースや物語を読んだ読み手の感情を予測する試みがある [Lin 08, Socher 11]．しかし，そうした研究とは異なり，対話という状況ではターゲットの過去の発話を利用することによって，予測精度の向上を期待できる．例えば，ターゲットの過去の発話から怒りの感情を抱いていたことを推定できれば，パートナーがなだめない限り，まだ怒りの感情を抱いている可能性が高い．我々はこの点に着目して，対話履歴に基づく新しい素性を提案し，その有効性を確かめる．

応答生成タスクにおいては，統計的機械翻訳を応用した発話応答生成の枠組み [Ritter 11] に基づく手法を提案する．感情タグ付きコーパスを利用することによって，特定の感情を喚起させることに特化したモデルの学習を行う．このとき，線形補間によるモデルの平滑化を行うことによって，疎データ問題に対しても頑健な応答生成を実現する．

実験では，6 億 4 千万発話から成る感情タグ付き対話コーパスを構築し，これを訓練データとして用いた．評価データには，両タスクにおいて共通して利用できるようなものを作業者 5 人が作成した．これらのデータを用いた実験の結果，提案手法の有効性を確認することができた．なお，後述するように，感情タグには [Plutchik 80] が定めた 8 つの基本感情を用いる．

本論文の構成は次のとおりである．2 章では，本研究で取り組む 2 つのタスクで訓練データとして用いる感情タグ付き対話コーパスの概要とその構築方法について述べる．3 章では，対話における聞き手の感情を予測する手法を提案する．4 章では，聞き手の感情を喚起する発話応答生成手法を提案する．5 章では，上記の 2 つのタスクに対する評価実験の結果について述べる．6 章では，関連研究について触れ，本研究との違いについて述べる．7 章では，全体のまとめと今後の課題について述べる．

2. 感情タグ付き対話コーパス

我々は，マイクロブログから大規模な感情タグ付き対話コーパスを自動構築し，これを訓練データとして利用する．このデータは，感情予測と応答生成の両タスクにおいて訓練データとして使われる．

本節ではその感情タグ付き対話コーパスの構築方法について述べる．はじめに，マイクロブログサービスの 1 つである Twitter から対話データを収集する方法を説明する．次に，手がかり表現を利用することで，収集した対話データ中の発話に対して，その発話者の感情を自動的にタグ付けする方法を説明する．

2·1 Twitter からの対話データの収集

対話データの収集は以下の手順で行う．

まず Twitter REST API[*2] を利用することにより発話 (ツイート) をクロールした．クロールされたデータは，2011 年 3 月から 2012 年 12 月までの期間に，77 万ユーザによって投稿された 55 億の発話で構成されている．そして，Twitter 特有の表現を削除，または変更することで，クロールされた発話をクリーニングする．

∗2 https://dev.twitter.com/docs/api/

表 1 Twitter から構築した対話データの統計情報

ユーザ数	672,937
対話数	311,541,839
発話の異なり数	1,007,403,858
対話数 / ユーザ数	463.0
発話数 / ユーザ数	1497.0
発話数 / 対話数	3.2

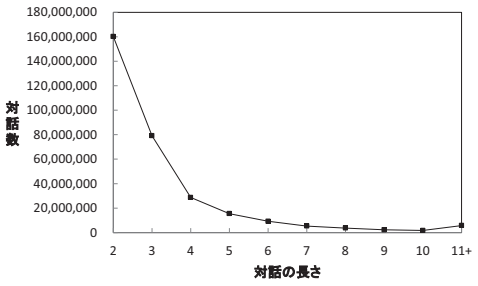


図 2 対話の長さの分布

- 引用を示すマークである RT や QT を含む発話を削除する．
- 発話中に URL があれば「URL」という文字列に置き換える．
- Twitter にはボット[*3]と呼ばれるユーザが存在する．ボットの発話をクロールしてしまうと，収集される発話内容に偏りが生じてしまう．そこで，ユーザ名に「bot」という文字列が含まれるユーザが投稿した発話を削除する．
- 発話が別ユーザへの返答となっているの場合，先頭または末尾にユーザ名が含まれる場合がある．そのような場合，先頭と末尾に現れるユーザ名は削除する．それ以外の位置にあるユーザ名は，それが発話者のユーザ名と一致すれば「私」，対話相手のユーザ名と一致すれば「あなた」，そのどちらでもない場合は「彼」に置き換える．

以上のような前処理を施したのちに，2 ユーザによる連続した発話のやりとり (対話と呼ぶ) を抽出する (表 4). 対話の同定には，Twitter REST API によって提供される ‘in_reply_to_status_id’ のフィールドを利用する．

表 1 に構築された対話データの統計情報を示す．また，図 2 に対話の長さの分布を示す．ここで対話の長さとは，その対話に含まれる発話総数である．ほとんどの対話 (98.2%) は高々10 発話で成り立っていた．また，最も長い対話は 1745 発話で構成されており，期間は 2012 年 1 月 2 日から 2 月 23 日までの間に渡っていた．

2·2 手がかり表現に基づく発話者の感情タグ付与

このようにして得られた対話データ中の各発話に対して，手がかり表現を利用して発話者の感情をタグ付けすることによって，表 4 のようなデータを作成する．これを感情タグ付き対話コーパスと呼ぶ．

*3 特定のキーワードに反応し自動的に応答するプログラム．

表 2 実験で用いた手がかり表現の一覧

感情	手がかり表現
怒り	イライラ，いらいら，腹が立つ，腹がたつ，はらが立つ，はらがたつ，腹立つ，はら立つ，はらたつ，ふざけるな，むかつくな，ムカくな，（怒），(｀へ´#)，(＃｀∧´)，(--メ)
期待	わくわく，ワクワク，楽しみ，期待大
嫌悪	嫌，嫌う，きらう，うんざり，ウンザリ，不快
恐れ	不安，心配，しんぱい，怖い，こわい
喜び	嬉しい，うれしい，幸せ，しあわせ，感激
悲しみ	悲しい，哀しい，かなしい，寂しい，さみしい
驚き	驚いた，おどろいた，びっくり，Σ（・□・;），！(◎_◎;)，Σ（゜д゜lll）
受容	安心，あんしん，頼りになる，ほっとする，頼もしい，たのもしい

感情のタグ付けにあたっては，Plutchik [Plutchik 80] が定めた基本感情である 8 カテゴリ (「怒り」「期待」「嫌悪」「恐れ」「喜び」「悲しみ」「驚き」「受容」) を採用した．タグ付けに用いた手がかり表現は，各感情と直接的な対応関係が見られ，なおかつ意味的な曖昧性がないと考えられるものを人手で選んだ．実際の実験で用いた手がかり表現の一覧を表 2 に示す．なお，手がかり表現を用いるという戦略上，タグ付けの漏れが発生することは避けられないが，後述するように感情タグ付き対話コーパスは各タスクの訓練データとして使うため，このことは大きな問題とはならない．

教師あり学習においてはタグ付けの精度が重要となるため，手がかり表現を含み，なおかつ，以下の 2 つの条件を満たす発話をタグ付与の対象とする．

(1) 手がかり表現が自立語を修飾していない
(2) 手がかり表現が否定，仮定，命令，疑問，譲歩，引用の表現を伴っていない

例えば「私は怒った父親が怖い」に「怒り」のタグは付与しない (手がかり表現「怒った」が自立語「父親」を修飾しているため，1 つ目の条件を満たさない)．また，「明日は遠足だから雨が降ったら悲しいなぁ」に「悲しみ」のタグは付与しない (仮定に基づく言明であるため 2 つ目の条件を満たさない)．2 つ目の条件の判定は，「ない」「ぬ」(否定)，「たら」(仮定)，「？」(疑問) などの語が，発話中に出現しているかをチェックすることにより行った．

表 3 に感情タグが付いた発話数とその精度を示す．タグ付けの精度は無作為に選んだ各感情カテゴリ 100 発話を 2 人の作業者によって調査した．具体的には，発話とそれに対して付与されたタグを提示し，そのタグ付与結果が妥当であるか否かを各作業者が判断した．2 人の判定結果の κ 係数は 0.85 となり，ほぼ完全な一致となった．表のタグ付け精度は 2 人の判定結果の平均である．ほぼ全ての感情カテゴリについて 95%を超える精度が得られている．ただし，ここでの評価手順は，(1) 自動タグ付け

表 3 タグ付け精度と感情がタグ付けされた発話数

感情	精度		発話数
	作業者 A	作業者 B	
怒り	95.0	95.0	197,756
期待	99.0	99.0	2,346,350
嫌悪	93.0	93.0	337,135
恐れ	96.0	96.0	2,671,222
喜び	94.0	96.0	2,247,105
悲しみ	97.0	97.0	533,931
驚き	97.0	97.0	830,372
受容	97.0	98.0	337,301

表 4 感情タグが付与された対話の例: 1 列目は発話 ID，2 列目は二人のユーザによる発話，3 列目は発話にタグ付けされた発話者の感情である．発話中の下線は，手がかり表現がマッチしたことを表す．

ID	発話	感情
1	**A**: 一緒に夕食にいかない？	
2	**B**: すみません．38 度の熱があるため行けません．	
3	**A**: そっか，寂しい な．早く良くなるといいね．	悲しみ
4	**B**: ありがとう．そういってくれて 嬉しい よ．	喜び

結果を作業者に提示している，(2) 全ての発話には特定の手がかり表現が含まれていることを考えると，精度が高く見積もられやすい設定になっている可能性があるため，数字の解釈には注意をされたい．別の評価方法としては，作業者にタグ付け結果を提示せず，8 種類の感情タグから自由に 1 つを選択してもらった結果と，自動付与されたタグの一致を見ることなどが考えられる．しかし，実際の発話には，どちらの感情とも取れるようなケースがありうる．そのため，そのような評価方法では，作業者間の一致率が低下することが予想され，信頼性のある数字を得ることが難しくなるだろうと判断し，現在のような評価方法を採用した．

3. 対話における聞き手の感情予測

本章では，聞き手の感情予測タスクを行う手法について述べる．このタスクにおける入力は，ターゲットの発話とそれに対するパートナーの応答である．そして，システムは，その応答の結果ターゲットが抱く感情を，Plutchik の 8 つの感情カテゴリから 1 つ選んで出力する．例えば，表 4 における最初の 2 発話が入力として与えられたとすると，出力となる感情カテゴリは「驚き」である．

このとき，ターゲットが複数の感情を同時に抱く可能性があるが，本論文では最も顕著な感情にのみ焦点を当て，通常の分類問題として上記のタスクを解くこととする．そして，一対他 (one-vs-rest) 法によって多値分類器の学習を行う．学習アルゴリズムとしては Passive Aggressive アルゴリズムを用いる．

素性には，ターゲットとパートナーの各発話から抽出した単語 n-gram を用いる ($n \leq 3$)．抽出された n-gram は特定の感情を喚起させるイベントや行為 (表 4 における「38 度の熱」)，発話のスタイルやトーン (表 4 における「ありがとう」) を捉えることができる．ターゲットの発話から得られた n-gram とパートナーの発話から得られた n-gram は別々の素性として扱う．

単語 n-gram だけでは素性ベクトルが疎になりやすいため，各発話から発話者 (ターゲットまたはパートナー) の感情を推定した結果も素性として用いた．話し手の感情は聞き手の感情に強く影響を与えることが報告されており [Kim 12]，パートナーの感情は有効な素性になると考えられる．また，ターゲットが直前の発話時にどのような感情を持っていたのかが分かれば，これも感情予測の有力な素性になると期待できる．こうした素性は，ニュースや物語などを読んだ読み手の感情を推定する場合には使うことが難しく [Lin 08, Socher 11]，対話における聞き手の感情に着目した本研究に特有のものである．

この素性を得るためには，2 章で述べた手がかり表現に基づく手法を利用して，発話から発話者の感情の推定を行う．もし発話中に手がかり表現が出現しなかった場合には，別途学習しておいた分類器を利用することで，発話者の感情を推定する．その分類器の学習には Passive Aggressive アルゴリズムを利用し，一対他法により多値分類器を構築する．訓練データの正例は，感情タグ付き対話コーパス中の感情タグが付与された発話を用いる．負例は，正例と同じ数の発話を他の感情カテゴリから無作為に選んだものを用いた．素性には，単語 n-gram を用いた．ただし，訓練データ中には必ず人手で用意した手がかり表現が含まれていることから，そのまま学習すると手がかり表現に大きな重みが与えられてしまうため，手がかり表現は素性から取り除いて学習した．

4. 聞き手の感情を喚起する発話応答生成

本節では，目的感情を喚起する応答生成手法について述べる．目的感情は，感情予測の場合と同じく，Plutchik [Plutchik 80] が定めた 8 つの基本感情カテゴリのいずれか 1 つとした．4·1 節では，Ritter ら [Ritter 11] が提案した統計的応答生成手法について述べる．4·2 節では，Ritter らの手法を拡張することによって，目的感情を喚起させるような応答を生成する方法を述べる．

4·1 統計的発話応答生成

我々は Ritter らの研究 [Ritter 11] と同様に，統計的機械翻訳の枠組みを利用することにより，与えられた発話に対して応答を生成する．この方法は，応答を入力発話に対する翻訳とみなして応答生成を行うというものである．そして，一般的な機械翻訳システムと同様，発話と応答のペアから機械翻訳ツールを利用することでモデルの学習を行う．

本研究では，フレーズベースの統計的翻訳手法 [Koehn 03] を用いる．日本語では，文を単語に分割する必要が

あるため，形態素解析器 MeCab を利用した．我々は，応答を生成するためのデコーダーには Moses [Koehn 07a] を利用した．

応答生成器を構成する翻訳モデルと言語モデルの学習には GIZA++ [Och 03] と SRILM [Stolcke 02] をそれぞれ利用した．このうち，翻訳モデルについては，後処理として次のように翻訳表をフィルタリングした．GIZA++ は対話データに直接適用されたとき，同じ単語を含むフレーズペアを学習しやすいため，そのまま利用するとオウム返しの応答が生成されやすいことが知られている [Ritter 11]．そこで，これを避けるために，フレーズ同士が部分文字列の関係にあるフレーズを翻訳表から取り除いた．また，機械翻訳の場合とは異なり，フレーズの配置は応答の適切さとは相関が強くないため，並べ替えモデルを用いない [Ritter 11]．

Moses における対数線形モデルの重みは，Moses が与えるデフォルトの重みを利用する．すなわち，翻訳モデルの各素性には 0.2，言語モデル素性には 0.5 の重みを与える．

4·2 感情に適応したモデルの利用

本節では，4·1 節の枠組みを用いて，目的感情に応じた応答生成を行う方法について説明する．

我々は，感情タグ付き対話コーパスを利用して，8 つの各感情を喚起することに特化した翻訳モデルと言語モデルを学習する．具体的には，各感情 e に対して，感情 e がタグ付けされた発話の直前の発話ペアを翻訳モデルの学習に利用する．例えば表 4 の対話が訓練コーパスとして与えられた場合,「悲しみ」の感情を喚起させる応答生成器の翻訳モデルの学習には，表 4 の発話 1 と発話 2 が用いられる．一方，言語モデルの学習には発話 2 のみが使われる．このようにして学習されたモデルを感情喚起モデルと呼ぶ．

しかしながら，感情喚起モデルだけを用いる方法では，データスパースネス問題の影響で上手く応答を生成することが難しい．なぜなら，感情タグ付き対話コーパスにおいては，全ての発話に感情がタグ付けされているわけではないため，学習に使うことができる発話数は全コーパスと比較すると少なくなってしまうからである．この問題に対処するために，感情を考慮せずに全コーパスで学習したモデルを学習し (これを一般モデルと呼ぶ)，翻訳モデルと言語モデルともに，感情喚起モデルと一般モデルの線形補間を行う．このような手法は，機械翻訳における分野適応において広く使われており，有効であることが知られている [Sennrich 12]．

翻訳モデルの線形補間には Moses 付属のスクリプト `tmcombine.py` [Sennrich 12]，言語モデルの線形補間には SRILM をそれぞれ利用した．翻訳モデルから得られる 4 つの素性 (2 つのフレーズ翻訳確率と 2 つの語彙重み) 全てに対しては，同じ重み α ($0.0 \leq \alpha \leq 1.0$) を利用した．一方，言語モデルに対する重みは β ($0.0 \leq \beta \leq 1.0$) とした．ここで重み α と β は，感情喚起モデルの強さを制御するパラメータであると解釈できる．$\alpha = 1.0$ (または $\beta = 1.0$) のときは，感情喚起モデルのみを用いることに相当する．逆に，$\alpha = 0.0$ (または $\beta = 0.0$) のときは，一般モデルのみを用いることに相当する．α と β が共に 0.0 のときは，4·1 節で述べたモデルと等しくなる．

表 5 評価データの例

	目的感情：喜び
発話:	16 歳になりました。これからもよろしくお願いします！
応答 1:	誕生日おめでとうございます！
応答 2:	おめでとう！今度誕生日プレゼントあげるね。
応答 3:	おめでとうー!!幸せな一年を！

5. 評 価 実 験

5·1 評 価 デ ー タ

我々は 5 人の作業者に評価データの作成を依頼した．表 5 に，作成された評価データの例を示す．一つの発話に対して，ある目的感情を喚起する応答が最大 3 つまで付与されているようなデータとなっている．この評価データは 2 つのタスク両方において利用する．ここで，一つの発話に複数の応答を用意しているのは，応答生成の評価に BLEU スコアを用いたためである (5·3 節参照)．

評価データの作成には，以下のような流れに沿って作業を行った．まず，各作業者に 80 発話ずつ与え，目的感情を喚起するような応答を作成してもらうよう依頼した．この 80 発話の内訳は，8 つの目的感情それぞれに対して 10 発話ずつとなっている．この 80 発話は，異なる作業者間で重複しないようにした．作業者の負担を減らすために，なお，感情タグ付き対話コーパスから無作為に抽出した発話を作業者に提示し，そこから応答を作成できそうな 80 発話を選択してもらっている[*4]．この作業の結果，400 (= 80×5) の発話応答ペアを得た．さらに，各発話に対して，5 人の作業者のうち別の 2 人にも新たな応答を作成してもらった．これによって，各発話に対して 3 つの応答が作成されることとなり，合計で 1200(=400×3) の発話応答ペアを得た．

最後に，評価データの質を確保するため，全ての発話応答ペアに対して，2 人の作業者が応答の適格性 (正しく目的感情を喚起できるかどうか) を調査した．このとき，作業者が自身の生成した応答の適格性判定に関与することのないよう，データの割り当てを行った．そして，もし両方の作業者が不適格だと見なした場合には，その応答を評価データから削除した．この結果，評価データに含まれる発話応答ペアの数は 1099 となった．なお，3 つ

*4 評価データ作成に使用した発話は訓練データから除去している．

表 6 分類器，翻訳モデル，言語モデルの学習に用いた発話ペア数

感情	発話ペア数
怒り	119,881
期待	1,416,847
嫌悪	333,972
恐れ	1,662,998
喜び	1,724,198
悲しみ	436,668
驚き	589,790
受容	228,974
合計	6,513,328

表 7 感情予測タスクの評価実験結果

感情	パートナー			パートナー & ターゲット		
	精度	再現率	F_1	精度	再現率	F_1
怒り	0.455	0.476	0.465	**0.600**	**0.548**	**0.476**
期待	0.518	0.526	0.526	**0.614**	**0.637**	**0.490**
嫌悪	0.275	**0.519**	0.359	**0.378**	0.511	**0.435**
恐れ	**0.484**	**0.727**	**0.581**	0.459	0.706	0.556
喜び	0.690	0.417	0.519	**0.720**	**0.590**	**0.649**
悲しみ	**0.711**	0.467	0.564	0.670	**0.562**	**0.611**
驚き	0.511	0.348	0.414	**0.584**	**0.437**	**0.500**
受容	**0.695**	0.452	0.548	0.682	**0.514**	**0.586**
平均	0.542	0.492	0.497	**0.588**	**0.563**	**0.567**

全ての応答が不適格と判断された事例はなかった．

5·2 感情予測タスクの結果

まず聞き手の感情予測に関する評価実験の結果を述べる．表 6 に各感情カテゴリごとの訓練データの数を示す．ここで言う訓練データとは，感情タグ付き対話コーパス中において，各感情タグを付与された発話の直前に出現する発話ペアのことである (例えば，表 4 における発話 1 と発話 2 は，感情カテゴリ「悲しみ」に対する訓練データとして使われる)．また，Passive Aggressive アルゴリズムの実装は opal[*5] を用いた．

対話を扱っている提案手法の特徴の 1 つは，パートナーの発話だけでなく，ターゲットの直前の発話からも素性抽出を行っている点である．そうした素性の効果を検証するため，以下の 2 種類の分類器を学習し，各感情ごとに精度，再現率，F_1 を比較した．

パートナー パートナーの発話中の単語 n-gram と，パートナーの感情推定結果の 2 種類の素性のみを用いた分類器．

パートナー & ターゲット 上記 2 種類の素性に加えて，ターゲットの発話中の単語 n-gram とターゲットの感情推定結果を素性として用いた分類器 (この分類器は，提案する素性を全て用いたものとなる)．

表 7 に 1099 の発話応答ペアに対する感情予測の結果を示す．パートナーの発話から作成した素性に加えて，ターゲットの直前の発話から作成した素性を利用することで，分類の精度および再現率が向上することが確認された．

表 8 に分類器 (表 7 のパートナー&ターゲット) の混同行列を示す．最も多く選んだカテゴリを太文字，最も多く間違えたカテゴリを下線で示している．この表からいくつかの典型的な混同する感情カテゴリのペアを見て取れる．分類器は「嫌悪」と「怒り」,「喜び」と「期待」,「恐れ」と「嫌悪」などを正しく区別することに失敗している．この原因は，これらの感情カテゴリが本質的に類似しているためである．実際に，Plutchik の分類では「怒り」と「嫌悪」,「期待」と「喜び」,「恐れ」と「嫌悪」がそれぞれ隣接しており，類似した感情とされている．

誤分類した対話を調査した結果，主に 2 種類の原因が見つかった．それぞれの対話の例を表 9 に示す．1 つ目は皮肉であり，これは語彙的な素性だけでは捉えることが難しいことが報告されている [González-Ibáñez 11]．2 つ目は，予測の判断材料となる情報の欠如である．ここでは「驚き」が正解となっているが，ターゲットが本当に「驚き」の感情を抱くかどうかは，パートナーの発話内容がターゲットにとって未知であるかどうかに強く依存する．しかし，そうした個人の背景知識に関する情報は現在の枠組みでは考慮できていない．

表 8 予測結果の混同行列．最も多く選んだカテゴリを**太文字**，最も多く間違えたカテゴリを 下線 で示している．

正解の感情 ＼ 予測した感情	怒り	期待	嫌悪	恐れ	喜び	悲しみ	驚き	受容	合計
怒り	**69**	0	26	20	0	8	2	1	126
期待	1	**86**	11	7	13	0	6	11	135
嫌悪	25	1	**68**	18	2	6	7	4	133
恐れ	3	0	22	**101**	1	5	9	2	143
喜び	1	28	9	4	**85**	1	7	9	144
悲しみ	6	3	25	14	5	**77**	5	2	137
驚き	7	10	9	32	5	7	**59**	6	135
受容	3	12	10	24	7	9	6	**75**	146
合計	115	140	180	220	118	115	101	110	1099

表 9 システムが誤った発話・応答ペアの例

発話:	後輩から告白ラッシュわず．バカにされてる！
	明日バッキバキにしてやるー（ンД´）ノ
応答:	はいはい，自慢乙．よかったですね〜。
	正解: 怒り，推定結果: 受容
発話:	黒髪がモテるってマジか。
応答:	80%くらいの男子は黒髪が好きらしい。
	正解: 驚き，推定結果: 恐れ

5·3 応答生成タスク

次に応答生成タスクに関する実験結果について報告する．感情喚起モデルは，表 6 にある発話ペアから学習した．一方，一般モデルは，感情タグ付き対話コーパスの 6 億 4 千万の発話ペアを利用した．ただし，計算機資源の制限から，一般モデルの翻訳モデルの学習には，その

*5 http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/

表 10 BLEU スコアの比較

一般モデル	0.64
提案手法	1.05
最適化された結果	**1.57**

うち約 400 万発話ペアのみを利用した．言語モデルの学習には 6 億 4 千万の応答全てを使って 5-gram モデルを学習した．

§1 自動評価

最初に，提案手法のベースとなっている [Ritter 11] と同様に，BLEU スコア [Papineni 02] による自動評価を行った．なお，[Ritter 11] においては，BLEU と人手評価にはある程度の一致が見られるものの，それだけでは十分な評価尺度たりえない可能性について言及が見られる．そのため，以下での結果は，あくまでも補足的なものであり，より良い自動評価方法の設計は今後の重要な研究課題の一つであると考えている．

この評価において，システムは評価データ中の発話と目的感情が与えられると，それに対する応答を生成する．そして，その生成結果と評価データ中の応答を用いて BLEU スコアを求める．重み α と β は 2 分割交差検定で調整した．α と β の値を $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ の中でそれぞれ変化させ，最も高い BLEU スコアが得られた値に決定した．なお，重みの値は感情ごとに異なる値を設定する．

3 つの手法による BLEU スコアの比較を表 10 に示した．表の 1 行目は一般モデルのみを用いた手法による結果である．これは提案手法において，重み α と β をともに 0.0 に設定した場合に相当する．2 行目は我々の提案手法による結果である．最後の行は全評価データに対する BLEU スコアが最大になるように α と β を調整した結果である．この結果が提案手法による BLEU スコアの上限値と考えることが出来る．

この自動評価の結果は，感情喚起モデルが聞き手の感情を喚起させる応答生成に有効であることを示している．BLEU スコアは 0.64 から 1.05 に上昇した．一方で，提案手法と最適化された結果にまだ差があるが，これは評価データが小さいために α と β が適切に調整できなかったことが原因であると考えられる．

§2 人手評価

上記で述べたように，BLEU スコアは客観的な評価尺度として有用性を認めることができるものの，その良し悪しだけで生成結果を議論することは危険であると我々は考えている．そこで，BLEU スコアによる自動評価に加えて，人手による評価を行った．この評価実験では，表 10 における一般モデル (以下，ベースラインと呼ぶ) と提案手法の 2 つを比較対象とした．なお，評価データにおける BLEU スコアに対して重み α と β を最適化した手法は，人手評価の作業コストを考慮して対象から除外した．

具体的な評価手順は以下の通りである．評価データ中の 400 発話に対して各手法によって応答を生成し，得られた合計 800 応答に対して 2 人の作業者がその適切さを評価した．応答は (1) 入力に対する応答として意味が通っている，(2) 目的感情を喚起している，という 2 つ基準を満たすならば適切であると判断した．前者の基準については，応答が日本語文として成り立っており，なおかつ，与えられた発話に対する適切な応答となっていた場合に，応答として意味が通っていると判断した．公平な比較を行うために，どちらのシステムが生成した結果であるのかは作業者には知らせなかった．また，2 つのシステムによる生成結果はランダムに提示した．

表 11 に人手評価の結果を示す．上記の 2 つの基準について，作業者のうちどちらか 1 人が適格であると判断した応答数は，ベースラインと提案法でそれぞれ 147 と 156 であり，提案法のほうが精度が高いことが確認された (表 11 の正解率)．さらに，本論文で提案した手法は基準 (2) を満たすような応答生成に関わるものであることから，基準 (2) について 2 人の作業者が共に適格であると判断した応答数の調査も行った．この結果，ベースラインと提案法では，それぞれ 74 と 92 の応答が適格であると判断された (表 11 の感情喚起率)．このことから，提案法は，より明確に感情を喚起するような応答生成を実現できていることが分かる．なお，評価作業の信頼性を確認するため，2 人の作業結果の一致度を調べたところ，κ 係数は 0.59 であり，中等度の一致を確認することができた．

2 つのシステムの差が統計的に有意なものであるかどうかを確認するため片側 t 検定を行った[*6]．その結果，正解率については有意差は見られなかったものの，感情喚起率については有意差が見られた (共に有意水準は 5%とした)．感情喚起率において有意差が見られたことから，提案手法によって，目的感情を喚起する応答が生成されやすくなっていることが確認できた．一方，正解率に有意差は見られなかったが，これは，基準 (1) に起因する誤りが多かったため，全体としてはシステム間に差がつきにくくなっているためであると考えられる．

今後，基準 (1) に起因する誤りを削減していくためには，提案手法のベースになっている Ritter らの手法を改善する必要があると考えられる．実際，統計的機械翻訳に基づく Ritter らの手法では，非文法的な文が生成されることを完全に防ぐことが難しく，そのために基準 (1) を満たすことができずに誤りと判定された事例が散見された．また，本手法では，オウム返しが過剰に生成されることを防ぐために翻訳表のフィルタリングを行ったが，適度にオウム返しを生成すること自体はむしろ自然なことであり，この点にも改善の余地があると思われる[*7]．

表 11 人手による評価の結果．正解率と感情喚起率の定義は本文を参照．括弧中の数字は適格と判断された応答の数．

	ベースライン	提案手法
正解率	0.37 (147)	0.39 (156)
感情喚起率	0.19 (74)	0.23 (92)

表 12 ベースラインと提案手法による応答生成の例

目的感情:	喜び
入力:	写真コンクールで優秀賞とったーーーー！ 2 月 7 日表彰式！
ベースライン:	残念ながら 1 日目だけですのー
提案手法:	金賞受賞おめでとう !!! おめでとうなのー
目的感情:	受容
入力:	自暴自棄になってるの？
ベースライン:	楽しみ！
提案手法:	きっとまだ大丈夫
目的感情:	期待
入力:	だよね！なんとかなるよね ww あ，わたしグッズ買わなきゃなのでその時間だけは取ってくれるとうれしい (´∀`)
ベースライン:	私はグッズ買ってないから不安ですね
提案手法:	いいですね！私も買いますね !!!

§3 議 論

表 12 にベースライン手法と提案手法による生成例を示す．最初の 2 つの例では，提案手法は目的感情 (「喜び」と「受容」) を喚起する応答を生成することができている．これらの例から「喜び」や「受容」の感情喚起モデルはそれぞれ祝福や安心させるフレーズに大きな確率を与えることができていると考えられる．また，最後の例では，話し手 (= システム) 自身が期待をしているような応答が生成されており，その結果として，聞き手にも期待を抱かせるようになっている．興味深いことに，これと同様の現象は，実際の会話においても観測されることが報告されている [Kim 12].

本論文では，Ritter ら [Ritter 11] の提案する統計的機械翻訳モデルに基づく応答生成を議論した．これは，Moses などの既存ツールを元に容易に実装可能な点を考慮してのことである．しかし，大規模な対話データから感情タグ付きコーパスを構築し，それを利用して感情を喚起するような応答生成を行うという考え方自体は，規則ベースや検索ベースなど，他の応答生成手法にも適用できると考えられる．例えば，規則ベースであれば，感情タグ付きコーパスから求めた何らかの統計値を規則に組み込むことが考えられる．また，検索ベースであれば，感情タグ付きコーパスにおいて，実際に目的感情を喚起している発話との類似性を考慮してスコア付けを行うことなどが考えられる．

*6 ベースラインは提案手法の特殊形であるため，提案手法の方が精度が高くなるはずであるという前提に基づいて片側検定を行った．

*7 現在の手法ではオウム返しが生成されにくくなっているものの，全く生成されないわけではないことに注意されたい．例えば「これからもよろしくお願いします」という入力に対して，「これからも → よろしくお願いします」と「よろしくお願いします → これからも」いうペアが学習されていれば，偶然ではあるが，上記の発話に対して「これからもよろしくお願いします」というオウム返しが生成されうる．

6. 関 連 研 究

6·1 感情予測に関連する研究

従来，感情に着目した研究としては，話し手 (または書き手) の感情に基づいて発話を分類する試みが多く行われている [Ayadi 11, Bandyopadhyay 11, Balahur 11]．これに対し我々は，話し手ではなく聞き手の感情の予測を行った．

テキストの読み手の感情を推定する研究については，Lin ら [Lin 08] や Socher ら [Socher 11] の研究がある．我々の研究は，これらの研究とは対話に焦点を当てている点で異なり，そのようなタスク設定では抽出できない対話履歴から作成した素性の有効性を検証した．

我々の研究は Tokuhisa ら [Tokuhisa 08] の研究とも関連が深い．この研究では，あるイベント (例: クリスマスプレゼントをもらう) によって喚起される感情 (例: 嬉しい) をウェブテキストから獲得している．こうした知識は，我々の感情予測タスクの精度向上などに有用であると考えられる．

対話における感情を扱った研究として，対話データにおける感情の移り変わりを調査した Kim ら [Kim 12] の研究がある．Kim らは話し手の感情を自動推定し，その移り変わりに関する定量的調査を行っている．

6·2 応答生成に関連する研究

対話における自動応答生成は長い歴史を持つ．従来の研究では，テンプレートを用いた手法が主流であったが，近年では大規模な対話データが手に入る環境が整ったことから，統計的なアプローチも試みられている [Ritter 11]．しかし，我々が知る限り，ユーザの感情をモデル化した統計的応答生成システムは提案されていない．

読み手の感情を喚起する文を生成する研究としては，ジョークやユーモアのあるテキストを生成する研究がある [Dybala 10, Labtov 12]．これらの研究は何らかの感情を聞き手に喚起するという点で我々の研究に似ているが，1 つの感情に特化している．これに対して，我々が提案する手法は 1 つの感情に特化しない．

翻訳モデルや言語モデルの線形補間は機械翻訳の分野適応として広く使われている [Koehn 07b, Sennrich 12]．しかしながら，応答生成において分野適応手法を用いた研究はなく，本研究は分野適応手法が応答生成においても有効であることを示した初めての研究となる．

7. お わ り に

本論文では対話における聞き手の感情予測と，聞き手の感情を喚起する応答生成を実現する方法を述べた．聞き手の感情予測では，対話履歴から素性を作成する手法を提案し，評価実験を通してその有効性を検証した．一方，応答生成においては，統計的機械翻訳を応用した応答生成の枠組みをベースとし，感情喚起モデルと一般モデルを線形補間によって組み合わせる手法を提案した．そして，自動評価と人手評価の両方により，提案手法の有効性を示した．

今後の課題としては，感情タグ付きコーパスをさらに大規模化することが挙げられる．本研究において構築した感情タグ付きコーパスは，650 万以上の発話ペアを含む大規模なものであり，この意味において提案手法は有効であったと言えるが，今後各タスクの精度を改善するためにはさらに大規模なコーパスが必要であると考えている．データ不足に起因する問題は，生成タスクで用いたモデル平滑化などのテクニックによってある程度は対処可能であるが，やはり本質的にはデータを大規模化することが必要であろう．また，大規模化と同時に，感情タグの自動付与の精度を高めることも重要である．例えば，提案手法では，発話単位で感情タグを付与しているが，文や句などのより詳細な単位にタグ付与を行い，どの部分が発話者の主たる感情なのかを判定することができれば，より精度の高いタグ付与につながることが期待できる．

予測と生成の両タスクにおいては，さらに長い履歴を持つ対話を使うことを今後検討していきたい．本論文では，発話と応答という 2 つの発話しか利用しなかったが，さらに長い履歴を使うことによって，処理精度を向上させることができるようになると期待できる．これ以外にも，特に統計的な応答生成は，現在においても十分に研究が進んでいない分野であるため，今後の発展が大いに期待できる．例えば，マイクロブログからはユーザの年齢，性別，職業といった属性情報が入手可能であるため，そのような属性を考慮した応答生成を行うことが考えられる．また，本研究では目的感情は入力の一部として与えられるとしていたが，これはアプリケーションによっては必ずしも妥当な仮定ではないと考えられる．そのため，例えば対話履歴から適切な目的感情を決定する方法などについて，今後見当を行う必要がある．

謝　　辞

本実験を遂行するにあたっては，東京大学の中島直哉氏，仁科俊晴氏，土屋圭氏，鈴木恵介氏，岡本大輝氏に協力して頂きました．ここに記して感謝いたします．

本論文は，Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics において発表した内容をもとに，議論を追加して再構成したものとなります [Hasegawa 13].

本研究は，最先端研究開発支援プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の支援を受けました．

◇ 参 考 文 献 ◇

[Ayadi 11] Ayadi, M. E., Kamel, M. S., and Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, Vol. 44, pp. 572–587 (2011)

[Balahur 11] Balahur, A., Boldrini, E., Montoyo, A., and Martinez-Barco, P. eds.: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics (2011)

[Bandyopadhyay 11] Bandyopadhyay, S. and Okumura, M. eds.: *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*, Asian Federation of Natural Language Processing (2011)

[Dybala 10] Dybala, P., Ptaszynski, M., Maciejewski, J., Takahashi, M., Rzepka, R., and Araki, K.: Multiagent system for joke generation: Humor and emotions combined in human-agent conversation, *Journal of Ambient Intelligence and Smart Environments*, Vol. 2, No. 1, pp. 31–48 (2010)

[González-Ibáñez 11] González-Ibáñez, R., Muresan, S., and Wacholder, N.: Identifying Sarcasm in Twitter: A Closer Look, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 581–586 (2011)

[Hasegawa 13] Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M.: Predicting and Eliciting Addressee's Emotion in Online Dialogue, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 964–972 (2013)

[Hasselgren 03] Hasselgren, J., Montnemery, E., Nugues, P., and Svensson, M.: HMS: A Predictive Text Entry Method Using Bigrams, in *Proceedings of EACL Workshop on Language Modeling for Text Entry Methods*, pp. 43–50 (2003)

[Kim 12] Kim, S., Bak, J., and Oh, A. H.: Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations, in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pp. 495–498 (2012)

[Koehn 03] Koehn, P., Och, F. J., and Marcu, D.: Statistical phrase-based translation, in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pp. 48–54 (2003)

[Koehn 07a] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180 (2007)

[Koehn 07b] Koehn, P. and Schroeder, J.: Experiments in domain adaptation for statistical machine translation, in *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 224–227 (2007)

[Labtov 12] Labtov, I. and Lipson, H.: Humor as Circuits in Semantic Networks, in *Proceedings of the 50th Annual Meeting on Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 150–155 (2012)

[Lin 08] Lin, K. and Hsin-Yihn, H.-H.: Ranking reader emotions using pairwise loss minimization and emotional distribution regression, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 136–144 (2008)

[Och 03] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003)

[Pang 12] Pang, B. and Ravi, S.: Revisiting the Predictability of Lan-

guage: Response Completion in Social Media, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1489–1499 (2012)

[Papineni 02] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318 (2002)

[Plutchik 80] Plutchik, R.: A General Psychoevolutionary Theory of Emotion, in *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pp. 3–33, New York: Academic (1980)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W. B.: Data-Driven Response Generation in Social Media, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 583–593 (2011)

[Sennrich 12] Sennrich, R.: Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation, in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 539–549 (2012)

[Socher 11] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D.: Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 151–161 (2011)

[Spertus 97] Spertus, E.: Smokey: Automatic Recognition of Hostile Messages, in *Proceedings of the Ninth Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence*, pp. 1058–1065 (1997)

[Stolcke 02] Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit, in *Proceedings of the 7th International Conference of Spoken Language Processing*, pp. 901–904 (2002)

[Tokuhisa 08] Tokuhisa, R., Inui, K., and Matsumoto, Y.: Emotion classification using massive examples extracted from the Web, in *Proceedings of the 22nd international conference on Computational Linguistics*, pp. 881–888 (2008)

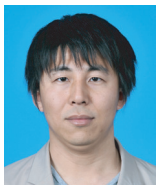
〔担当委員：東中 竜一郎〕

2013 年 4 月 30 日　受理

著　者　紹　介

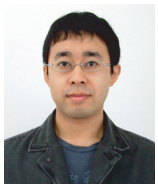
長谷川　貴之

2011 東京工業大学工学部情報工学科卒，2013 東京大学大学院情報理工学系研究科修士課程修了．修士 (情報理工学)．



鍛治　伸裕

2005 年 東京大学情報理工学系研究科博士後期課程修了．博士 (情報理工学)．同年 東京大学生産技術研究所産学官連携研究員，特任助手，特任助教を経て，現在，同大学生産技術研究所特任准教授．自然言語処理に興味を持つ．



吉永　直樹

2000 東大・理学部情報科学科卒業．2002 同大学大学院理学系研究科修士課程修了．2005 同大学大学院情報理工学系研究科博士課程修了．博士 (情報理工学)．2002 より 2008 まで日本学術振興会特別研究員 (DC1, PD)．2008 東大・生産技術研究所特任研究員，特任助教を経て 現在，同大学生産技術研究所特任准教授．計算言語学・機械学習の研究に従事．

豊田　正史

1994 東京工業大学理学部情報科学科卒．1996 同大学大学院情報理工学研究科修士課程修了．1999 同大学大学院情報理工学研究科博士後期課程修了．博士 (理学)．同年，科学技術振興事業団計算科学技術研究員．2001 東大・生産技術研究所学術研究支援員，同大学同研究所産学官連携研究員，同大学生産技術研究所特任助教授，助教授を経て　現在，同大学生産技術研究所准教授．ウェブマイニング，情報可視化に興味を持つ．