社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

聞き手の感情を喚起する発話の分類と生成

長谷川貴之† 　鍜治　伸裕†† 　吉永　直樹†† 　豊田　正史††

† 東京大学大学院情報理工学系研究科　〒 113–8654 東京都文京区本郷 7–3–1
†† 東京大学生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †{hasegawa,kaji,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

あらまし　聞き手の感情を推し量ることは，自然な対話を行うためには必要不可欠な作業である．本論文では，聞き手の感情を考慮した発話生成の実現に向けた足がかりとして，(1) 聞き手にどのような感情を喚起させるのかという観点に基づく発話分類，(2) 聞き手に特定の感情を喚起させるような発話生成，という 2 つのタスクを提案する．加えて，いずれのタスクに対しても，マイクロブログ上の大規模対話データを用意し，少数の感情表現で対話に聞き手の感情カテゴリをタグ付けした．そのコーパスに基づく統計的手法を提案し, 実験を通して有効性を示す.

キーワード　対話，感情，発話生成

Classification and Generation of the Utterances Evoking Receiver's Emotions

TAKAYUKI HASEGAWA†, NOBUHIRO KAJI††, NAOKI YOSHINAGA††,
and MASASHI TOYODA††

† Graduate School of Information Science and Technology, the University of Tokyo　Hongo 7–3–1, Bnkyo-ku, Tokyo, 113–8654 Japan
†† Institute of Industrial Science, the University of Tokyo　Komaba 4–6–1, Meguro-ku, 153–8505 Japan
E-mail: †{hasegawa,kaji,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract　Considering receiver's emotion in conversation is important to do natural communication. We propose two tasks. (1)Classification of utterances in terms of receiver's emotion. (2)Generation of utterances to evoke particular emotion. We prepare the conversation corpus of microblog, and tag it receiver's emotions by a few emotional phrases. The experiments show our statistical methods are better than baselines.

Key words　conversation, emotion, utterance generation

1. はじめに

我々は常日頃から相手の感情を意識しながら発話を行っている．例えば，落ち込んでいる相手を鼓舞する，または，相手の感情を害するような発言を避ける，などといったことは誰もが日常的に行っていることであろう．そのため，計算機による発話生成において，聞き手の感情をモデルに組み入れることは自然な発話を生成する上で重要である．しかしながら，発話生成という分野自身が依然として未成熟であるという事情もあり，発話生成において聞き手の感情をどのようにモデル化すべきかという問題に対して，これまでに十分な知見の蓄積は行われていない．

本論文では，聞き手の感情を考慮した発話生成の実現に向けた足がかりとして，(1) 聞き手にどのような感情を喚起させるのかという観点に基づく発話分類，(2) 聞き手に特定の感情を喚起させるような発話応答生成，という 2 つのタスクを提案し，それらを実現するための方法を議論する．これら 2 つのタスクは，発話生成システムのモジュールとして必要になると予想されるだけでなく，直接的な応用を期待することもできる．まず前者の分類タスクは，例えば聞き手が気分を害するような発話を検出し，書き手に対して警告を発するようなシステムへの応用が期待される．そうしたシステムは，例えばオンラインコミュニケーションやカスタマーサービスセンターなどにおけてニーズを予想することができる．一方，後者の発話応答生成も，オンラインコミュニケーション環境における予測入力システムの開発につながると考えられる．

我々は上記の 2 つのタスクに対して統計的手法の適用を試みる．統計的手法は，自然言語処理の分野において大きな成功を

収めているものの，大規模な対話データは依然として取得困難であり，統計的な発話生成に関する研究事例は極めて数が少ない．しかしながら，近年では，CGM の普及に伴い，ウェブ上に大量の対話データが蓄積されつつあり，統計的学習手法の適用可能性が広がりつつある．そうしたウェブ上の対話データの 1 つである Twitter に対して半自動的に感情タグを付与することにより，統計的学習に必要となる大規模な訓練データの構築を行った．

以下，本稿の構成を述べる．本節は 1 節である．2 節では，関連研究について述べる．3 節では，コーパスに関する定義やタクスの説明をする．4 節では，感情対話コーパスについて述べる．5 節では，聞き手が喚起された感情を分類するタスクについて述べる．6 節では，聞き手に特定の感情を喚起させる発話を生成するタスクについて述べる．7 節では，5 節，6 節の実験結果に基づいて感情を喚起させる発話について考察する．そして 8 節にて本稿をまとめる．

2. 関 連 研 究

2.1 感 情 分 類

従来，感情に着目した研究としては，書き手/話し手の感情にもとづいてテキスト/発話を分類する試みが多い [1, 2]．これに対し，我々は，書き手/話し手ではなく，その読み手/聞き手の感情に基づく分類を試みている．

テキストの書き手の感情ではなく，読み手の感情を推定する研究については，Lin ら [3] や Socher ら [4] の研究がある．また，Tokuhisa ら [5] 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．これらの研究は，我々が行う分類タスクと問題設定が近いが，対話における感情を扱っているわけではない．

Kim ら [6] は対話データにおける感情の移り変わりを調査した．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．

2.2 発話の生成

発話生成は chatterbots の研究とともに長い歴史を持つ．特に近年においては，大量の対話データがウェブから取得可能になったことを受けて，そうしたデータに基づく統計的手法が盛んに研究されている．Ritter ら [7] は，統計的機械翻訳の手法を応用することにより，Twitter のデータから発話に対する応答の生成を試みている．130 万の対話ペアを用意し，機械翻訳と同じように対話テキストをパラレルコーパスと見なして利用した．Pang ら [8] は，モバイル環境における予測入力システムを開発した．ユーザーが受け取ったテキストと入力途中の返信テキストを素性として学習した．

聞き手の感情を考慮した発話生成研究という点では，ジョークの生成 [9] が我々の研究に最も近い．本研究では，ジョークという発話を生成するのではなく，特定の感情に特化しない汎用的な方法で，様々な感情を喚起する発話を生成する．

3. タスクの説明

本節では，本研究で扱う 2 つのタスクについて具体例を交えて説明する．

3.1 聞き手の感情に基づく発話分類

聞き手の内面に喚起される感情に基づいて，発話を分類するというタスクを考える．

本研究で利用する感情のカテゴリは，Plutchik[10] に従い，怒り (anger)，期待 (anticipation)，嫌悪 (disgust)，恐れ (fear)，喜び (joy)，悲しみ (sadness)，驚き (surprise)，受容 (trust) の 8 カテゴリを考える．以下に，発話とそれが喚起する感情の例を示す．この例では「喜び」を喚起している．

ユーザ A 大人かわいい♪ショート似合いますね (^o^ゞ
ユーザ B わぁ，ありがとぉぉお嬉しい (*^^*)

このように，2 つの連続した発話を対話ペアと呼ぶ．また，応答される発話 (ユーザ A の発言) を status，応答する発話 (ユーザ B の発言) を response と呼ぶ．

一般に，ある発話が喚起する感情は一意に定まるとは限らず，複数の感情を同時に喚起することも想定される．以下の例では，聞き手は「驚き」と「恐れ」を同時に喚起されている．

ユーザ A ぶっちゃけ，ストーカー機能みたいなもんだろ
ユーザ B その機能，私もびっくりしましたよー．怖い怖い．

そこで我々は，本タスクを単純な分類問題ではなく，マルチラベル分類問題と捉える．

3.2 聞き手の感情を考慮した応答生成

2 つめのタスクとして，Ritter らにならって，ある発話が与えられたときに，それに対する応答を生成するというタスクを考える．しかし Ritter らとは異なり，8 つの感情カテゴリの 1 つも入力として与えられるとし，聞き手にその感情を喚起するような応答を生成するという問題を考える．以下の例のように，ある発話 (ユーザ A の発言) に対して，ある特定の感情 (喜び) を喚起する応答 (ユーザ B の発言) を生成することを目標とする．

ユーザ A う〜熱が下がらない…
ユーザ B 調子はどんな感じすか？何か要る？買ってきますよん (^ ^)

もしユーザ B の発言を「もしかしたらインフルエンザかもよ…」とすれば，ユーザ A は「恐れ」の感情を持った応答をするだろう．このように聞き手の感情によって応答を変化させる．

4. 感情タグ付き対話コーパス

本節では，2 つのタスク行うシステムの学習および評価に用いるコーパスについて述べる．

4.1 少数感情表現による感情タグ付き対話コーパスの作成

response に感情がある対話において，status は感情を喚起する発話である可能性が高い．そこで，本研究で扱う 2 つのタスクでは，response の感情がタグ付けされた感情タグ付き対話コーパスを利用する．このコーパスを教師データとして，response の感情を分類する分類手法，response の感情を喚起する応答生成手法を学習する．このコーパスは，Twitter から取得した対話ペアに対して，人手で作成した少数の規則を適用することによって自動構築する．なぜなら，人手で訓練データを作成するとコストが大きく，大規模なデータを扱うことができないからである．その一方で，機械学習を用いて分類する方法も考えられるものの，8 カテゴリや 9 カテゴリの分類精度が 90% 以上は期待できない．人手で作成した規則を適用する場合は，データ量が少なくなってしまう反面，高い精度を期待できる．さらに，元々の対話ペアが増えれば増えるほど，タグ付けできる対話は増えていく．

ツイートデータは，Twitter API を用いて収集した 2011/3 〜 2012/5 までのツイートを利用する．このツイートデータには約 3.8 億の日本語の対話ペアがある．この約 3.8 億の対話ペアから response に感情を持つ対話ペアを判定し，感情タグ付き対話コーパスを作成する．

前処理として，ツイート中に「RT」や「QT」のような引用を示す Twitter 独自のマークが現れる対話ペアは削除した．引用されている文が感情を持つケースが考えられるほか，なるべく本来の対話に近づけたいからである．

このようにして得た対話ペアに対して，response の話し手の感情を，人手で作成した少数の感情表現をもとに推定することにより，感情タグ付き対話コーパスを作成する．各感情カテゴリに対応した感情表現を人手で作成し，もし response に感情表現を含む場合，その感情表現のカテゴリを response の話し手の感情としてラベル付けする．作成した感情表現は以下の通りである．各感情表現は，ひらがな，カタカナ表記も含めている．例えば,「嬉しい」であれば「うれしい」も含む.

怒り　いらいら，腹が立つ，ふざけるな，むかつくな，(怒)，(｀へ´＃)，(＃｀∧´)，(--メ)
期待　わくわく，期待，楽しみ
嫌悪　不快，うんざり，嫌う，嫌
恐れ　怖い，心配，不安，((((；Д))))
喜び　嬉しい，幸せ，感激
悲しみ　悲しい，哀しい，寂しい
驚き　驚いた，びっくり，えっ，Σ（・□・;），!(◎_◎;)，Σ（д lll)
受容　安心，頼りに，ほっとする，頼もしい

しかし，感情表現と一致しただけの対話には，意図した感情とは異なる感情を持つ対話ペアが多く含まれる．精度を高めるために，response と感情表現をマッチさせるときには以下のような表現が response に含まれる対話を削除する．

感情	怒り	期待	嫌悪	恐れ	喜び	悲しみ	驚き	受容
ペア (万)	8.1	137.9	16.3	99.9	117.2	25.5	42.8	187.8

表 1　各感情カテゴリの対話ペア数

感情	怒り	期待	嫌悪	恐れ	喜び	悲しみ	驚き	受容
精度 (%)	94	99	97	96	95	96	97	92

表 2　ランダムサンプリングした感情タグ付き対話ペアの精度

仮定形　たら，なら
(例) 雨が降ったら嫌だな
引用　と，とか
(例) 腹が立つと言ってた人がいた
否定形　ない，ぬ
(例) 期待してない
命令形　しろ，ください
(例) 安心してください
逆接　が，けど
(例) 最初はイライラしたけど

これらに加え，自立語が感情表現より後ろに現れる場合も削除する．なぜなら,「イライラした父親」のように，感情表現が自立語に対する修飾語になってしまうことが多いからである．ここでは，発話の最も後ろに出現する自立語が主要部分であると仮定する．

4.2 コーパスの精度とサイズ

4.1 節で説明した手法によってコーパスを構築した結果，表 1 のデータサイズとなった．最も少ないカテゴリが「怒り」であり約 8 万ペア，最も多いカテゴリが「期待」で約 128 万ペアだった．

また，各カテゴリごとランダムで選んだ対話 100 ペアずつを人手で調べたときの精度は，表 2 のようになった．このように，平均 95% 以上の精度とすることができた．

5. 聞き手の感情にもとづく発話分類

本節では，ある発話内容が与えられたとき，その聞き手の感情を分類するというタスクを提案し，SVM (support vector machine) を用いた手法およびその評価実験の結果を示す.

5.1 提 案 手 法

2 値分類器である SVM を利用した手法を提案する．8 つの感情カテゴリについて，8 つの SVM 分類器を学習する．発話が与えられたときは，各分類器を順に適用することによりラベルを割り当てる．このとき複数の感情カテゴリが割り当てられることもある．もし，8 つのどの感情カテゴリも割り当てられなかった場合は，感情無しとする．

提案手法の素性は，単語 n-gram と status の感情とした．単語 n-gram を素性とすることで，感情を喚起する出来事や口調を捉えることが期待できる．この実験では，単語 unigram，単語 bigram，単語 trigram まで素性として加えている．また，

	emotion			n-gram			n-gram+emotion			human		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
怒り	0.112	0.820	0.197	0.556	0.700	0.619	**0.588**	**0.940**	**0.723**	0.595	0.500	0.543
期待	0.157	**0.440**	0.231	0.545	0.360	0.434	**0.549**	0.390	**0.456**	0.608	0.790	0.687
嫌悪	0.136	**0.320**	0.190	0.280	0.260	0.269	**0.413**	0.310	**0.354**	0.500	0.430	0.462
恐れ	0.134	**0.350**	0.194	0.404	0.230	0.293	**0.472**	0.250	**0.327**	0.518	0.430	0.470
喜び	0.138	**0.380**	0.202	0.369	0.240	**0.291**	**0.359**	0.190	0.248	0.447	0.510	0.477
悲しみ	0.119	0.240	0.159	0.405	0.340	0.370	**0.427**	**0.350**	**0.385**	0.672	0.410	0.509
驚き	0.134	**0.360**	**0.195**	0.190	0.120	0.147	**0.191**	0.120	0.147	0.269	0.250	0.259
受容	0.117	0.250	0.159	0.324	0.220	0.262	**0.424**	**0.280**	**0.337**	0.610	0.360	0.453
感情なし	**0.200**	0.090	0.124	0.143	**0.600**	0.230	0.164	0.590	**0.257**	0.258	0.490	0.338
マクロ平均	0.138	0.361	0.184	0.357	0.341	0.324	**0.398**	**0.380**	**0.359**	0.497	0.463	0.467

表 3　各分類手法による各感情カテゴリごとの実験結果

status の感情を素性として加える．Kim ら [6] が述べているように，聞き手と話し手の感情は一致しやすいため，単語 n-gram の他に感情を素性として加えることで分類精度が向上する可能性がある．status の感情は，SVM と 4.1 節で感情推定に用いたルールを組み合わせて求める．もし status に感情表現が含まれる場合は，status の感情は感情表現に対応する感情カテゴリとする．感情表現が status にない場合は，感情表現で集めた対話ペアの response から学習した SVM を適用する．この SVM は，response から感情表現を含まない単語 n-gram を素性として学習する．response には必ず感情表現が含まれるので，感情表現を取り除かないと不適切な重みが感情表現についてしまい，正しく学習できない．

5.2 評 価 実 験

5.1 節で提案した手法の有効性を示すために評価実験を行った．4 節で説明したコーパスから正しいラベルがつけられている対話ペアを，各カテゴリに対し 100 件ずつ人手で取り出し，これを評価データとした．訓練データは，それ以外の対話ペアを用いるが，各カテゴリにおけるテストデータのサイズが均一であるため，訓練データもサイズを揃えるために対話ペアが最も少ないカテゴリである「怒り」に合わせ，全てのカテゴリで 8 万 500 ペアを利用した．

評価する手法は，emotion，n-gram，n-gram+emotion，human である．提案手法 (n-gram+emotion) は，5.1 節で述べた通りである．ベースラインとして，n-gram と，status の感情を推定し，その感情を聞き手の感情とする手法 (emotion) 選んだ．status の感情を推定する方法には，提案手法と同じように求める．また，人手でラベルをつけた結果 (human) も示す．これは，第 1 著者が status だけを見て response の話者の感情を推定し，各対話に 1 つのラベルをつけたときの結果である．

評価実験の結果は，表 3 のようになった．各手法の精度，再現率，F 値は 9 カテゴリのマクロ平均値となっている．実験結果から，単語 n-gram に status の話者の感情を加えた n-gram+emotion が他の手法よりも F 値が高いことから提案手法の有効性が確認できる．約 35 %は正しく分類できていることが分かる．しかし，人手の結果を見てみると，約 50 % とあまり高いとは言えない．これは聞き手の感情を正確に判定することは人間にも難しいことを示している．

感情のカテゴリに注目すると,「怒り」や「嫌悪」が高い精度で分類できている一方で,「驚き」はうまく分類できていない．「驚き」の場合は，他の感情と共起することが多く曖昧になりやすいことに加え，人が驚く出来事を学習できていないと正しく分類できない．以下の例は，正解が「驚き」であるのに，誤って「喜び」と判断した status である．

status　そういえば、今日は自分の鼻歌で目が覚めた。

また，分類された対話を見ると，話し手と聞き手の感情が一致する対話では正しく感情を分類できることが多かった．話し手の感情を素性に加えたことにより精度が向上している．以下の例は，評価実験において正しく分類できた例であり，status, response の話し手は共に嫌悪感を抱いている．

status　これ程窓ガラスが曇った電車なんて乗ったこと無いさな…
response　蒸す日は嫌です！;w;

6. 聞き手の感情を考慮した応答生成

次に，聞き手に特定の感情を喚起させるような応答を生成する手法および評価実験の結果を示す．

6.1 提 案 手 法

Ritter ら [7] が提案した応答生成手法を利用する．Ritter らは，フレーズベースの統計的機械翻訳手法を応答生成に応用した．統計的機械翻訳手法では，翻訳モデルと言語モデルの積で翻訳された文の正しさをモデル化する．翻訳モデルでは翻訳としての正しさを，言語モデルでは文法の正しさや流暢さをモデル化している．原言語を f，目的言語を e としたとき，翻訳モデルを $P(f|e)$，言語モデルを $P(e)$ とすると，$\arg\max_e P(f|e)P(e)$ となる目的言語テキスト e を求めることによって翻訳を行う．ここで，この原言語を status，目的言語を response とみなすことによって，応答生成を行うことを提案した．$P(f|e)$ は status と response の 130 万ペアから IBM モデル [10] でモデル化し，言語モデルは対話ペアの全ての response から n-gram モデルで学習した．

	怒り	期待	嫌悪	恐れ	喜び	悲しみ	驚き	受容	平均
emotion(α=0.9)	0.00	**0.79**	**0.58**	0.54	0.54	**0.45**	**0.51**	**0.71**	**0.52**
normal(α=1.0)	0.00	0.77	0.00	**0.56**	**0.58**	0.41	0.00	0.00	0.29

表 4　各分類手法による各感情カテゴリごとの実験結果 (BLEU)

私たちが提案する手法では，言語モデルの学習に工夫を行う．対話コーパスにおける全 response から学習した言語モデルと，感情対話コーパスにおいて特定の感情がアノテートされた対話ペアの status から学習された言語モデルの線形補間を用いる．以下に線形補間を行った言語モデル $P(e)$ の式を示す．対話コーパスにおける全 response から学習した言語モデルを $P_{normal}(e)$，特定の感情がアノテートされた対話における status から学習した言語モデルを $P_{emotion}(e)$ で表し，それぞれの言語モデルの重みは α とした．

$$P(e) = \alpha P_{normal}(e) + (1-\alpha)P_{emotion}(e) \quad (1)$$

これにより，特定の感情を喚起しやすい発話の生成確率が向上すると期待できる．MERT [11] により，システムのパラメータをチューニングできるが，予備実験では良い発話を生成できていなかったので今回はチューニングしなかった．

Ritter らは，SMT モデルを発話生成に適用したときの問題点を指摘し，それらに対する解決策を提案している．本研究でも ritter らにならって，SMT モデル学習の後，以下の後処理を行う．まず，以下の例のようにオウム返しのように受け取った文と同じ単語を復唱することを防ぐために，モデル学習後に部分文字列となっているフレーズペアを削除した．

status　国民が知らないと思っ て嘘の説明のオンパレード！
response　知らない国民に嘘のオンパレード！

さらに，Ritter らはフィッシャーの正確確率検定により，不要なフレーズペアを削減していたが，本研究では低頻度な言葉をなるべく発話しないようにするため，フレーズの頻度が 5 回以下のフレーズは削減した．

6.2 評価実験

評価は機械翻訳の評価尺度である BLEU を用いた．式 (1) の線形補間した言語モデルを使ったときと使わないときではどの程度の差がでるのか検証した．学習には機械翻訳用ライブラリ Moses(注1) [12] を利用した．言語モデルの学習は IRSTLM(注2) [13] を利用し，翻訳モデルの学習には IBM モデルを実装した GIZA++(注3) [14] を利用した．言語モデルは 3-gram までと設定した．

テストデータは，連続した 3 つのツイートからなる対話のうち最後のツイートが感情を含んでいる対話を用いる．これを各カテゴリに対して，100 対話用意した．

以下の実験で用いる翻訳モデルは，約 200 万対話ペアで学習を行った．P_{normal} は，2 億の response で学習した．また，$P_{emotion}$ は，各感情カテゴリにおいて，テストデータに含まれない感情タグ付き対話コーパスの status 全てで学習した．式 (1) における α の値を変化させて実験した．

α が 0.9，1.0 のときの実験結果を表 4 に示す．表の数字は，生成した応答とテストデータの応答から計算した BLEU 値である．式 (1) の線形補間した言語モデル (emotion) を使った手法で生成した応答の平均値は 0.59，使わなかった手法 (normal) は 0.29 となった．BLEU 値は α が 0.9 のとき最も大きく，α を小さくしていくにつれて小さくなっていった．「期待」「恐れ」「喜び」のカテゴリにおいては，α の値が大きくなっても BLEU 値が下がらなかったことから，結果にはデータサイズが大きく関係しているのではないかと考えている．今後は，「怒り」のようなデータサイズが小さいカテゴリでは，3-gram ではなく，1-gram を用いるようなスパースになりにくい方法を検討している．

7. 考　　察

a）感情タグ付き対話コーパスの質

まず分類タスク，生成タスクで共に利用した感情タグ付き対話コーパスの質について考察する．本研究で収集した会話ペアは，対話の一部を切り取っているため，status が誰かに向けられた返事であったときに誤りが含まれうる．例えば，悲しみを訴える人に対して慰める status に対し，継続して悲しみを訴える response を対話ペアとして切り取った場合，聞き手は話し手の発言 (status) に悲しみの感情を喚起されてないため，聞き手の感情を喚起する発話とは異なり不適切である．

このような対話をコーパスから削除するには，対話の先頭のみを status とみなすという方法が考えられる．しかしながら対話コーパスを対話の先頭のみから収集すると，対話コーパスの量が大きく減ってしまうため(注4)，分類タスクにおいてはかえって分類精度が低下する結果となった．今後は，3 つのツイートで成り立つ対話の最初と最後がツイートが同じ感情であるときだけデータから削除するといった方法で誤りを削除するつもりである．なぜなら，最初と最後のツイートの感情が同じであるとき，その間のツイートは聞き手の感情を喚起していないと考えられるからである．

b）喚起される感情の曖昧性

今回，聞き手の感情を喚起する発話分類では，対話コーパスからランダムに選んだ対話ペアをテストデータに用いたが，実験の結果を見てもわかるように人間でも判断が難しいケースが

(注1)：http://www.statmt.org/moses/
(注2)：http://hlt.fbk.eu/en/irstlm
(注3)：http://code.google.com/p/giza-pp/
(注4)：予備実験では約 1/3 になった

多数存在した．これは発話から喚起される感情には本質的に曖昧性がある [3] ためであり，例えば，「雪が降るらしいよ」という発話に対しては，喜ぶ人もいれば不快に感じる人もいる．このような喚起する感情に曖昧性がある発話をどのように扱うは難しいところであるが，まずは，複数の被験者によりタグ付けを行い，その一致度が高い「特定の感情を強く喚起する発話」をテストデータとして用いて評価を行うことを検討している．

c） 評価方法の検討

応答生成では，先行研究 [7, 8] でも述べているように高い BLUE 値はでなかった．確かに，応答生成の結果を見ると，正しい日本語になっている発話もあるものの，ほとんどの応答では正しい日本語になっていない．

しかし，本質的に出力に曖昧性のある応答生成というタスクの評価に機械翻訳の評価尺度である BLEU を用いることの是非も再検討する余地があるだろう．この点については，被験者に生成した発話を与えて（このような応答がありうるか）人手で直接評価してもらうことを検討している．また一方で，上で述べた分類タスクにおけるテストデータの生成方法と同様に，人間が返答を予想しやすい対話ペアをテストデータとすれば，BLUE を用いた場合でも妥当な評価が行えるのではないかと考えている．

8. おわりに

本稿では，発話によって喚起される聞き手の感情を分類するタスクと，特定の感情を喚起させるための発話生成タスクを提案した．どちらのタスクにおいても感情を考慮した手法を用いることにより，精度を向上させることができた．今後の課題は，コーパスの精度とテストデータの質の向上，そして評価方法の検討である．

文　献

[1] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 806–814, Stroudsburg, PA, USA, 2010.

[2] Das D. and Bandyopadhyay S. Analyzing emotional statements ? roles of general and physiological variables (saaip 2011). In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*, pp. 59–67, Chiang Mai, Thailand, November 2011.

[3] Kevin Lin and Hsin-Hsi Hsin-Yihn. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 136–144, 2008.

[4] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 151–161, 2011.

[5] Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1 of *COLING '08*, pp. 881–888, 2008.

[6] Suin Kim, JinYeong Bak, and Alice Haeyun Oh. Do you feel what i feel? social aspects of emotions in twitter conversations. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM '12, pp. 495–498, 2012.

[7] Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 583–593, 2011.

[8] Bo Pang and Sujith Ravi. Revisiting the predictability of language: Response completion in social media. In *EMNLP-CoNLL*, pp. 1489–1499, 2012.

[9] Pawel Dybala, Michal Ptaszynski, Jacek Maciejewski, Mizuki Takahashi, Rafal Rzepka, and Kenji Araki. 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.

[10] F. Brown Peter, A. Della Pietra Stephen, J. Della Pietra Vincent, and L. Mercer Robert. The mathematics of statistical machine translation: Parameter estimation. *Comput.Linguist.*, Vol. 19, No. 2, pp. 263–311, June 1993.

[11] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pp. 160–167, 2003.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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, Prague, Czech Republic, June 2007.

[13] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. Irstlm: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pp. 1618–1621, 2008.

[14] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, Vol. 29, No. 1, pp. 19–51, March 2003.