2L4-OS-07a-5

用例ベース対話システムにおける用例の評価値推定

Satisfaction Estimation for Examples in Example-based Dialogue Systems

水上 雅博 *1 Masahiro Mizukami
Lasguido Nio*1 Lasguido Nio
木付 英士 *2 Hideki Kizuki
野村 敏男 *2 Toshio Nomura
Graham Neubig*1 Graham Neubig
Sakriani Sakti*1 Sakriani Sakti
戸田 智基 *1 Tomoki Toda
中村 哲 *1 Satoshi Nakamura

*1奈良先端科学技術大学院大学 Nara Institute of Science and Technology
*2シャープ株式会社 SHARP Corporation

In dialogue systems, dialogue modeling is one of the most important factors contributing to user satisfaction. Especially in example-based dialogue modeling, effective methods to build and evaluate dialogue example database are the key to dialogue quality. However, it is difficult to build a high-quality example database. In this paper, we propose a model predict how users will evaluate examples in example-based dialogue systems. This prediction model estimates the prospective evaluation score of unknown examples from already-known examples making it possible to evaluate the quality of examples without subjective evaluation or dialogue experiments. Further, this example prediction model we perform experiments using this model to select the utterance used to respond to the user. Experimental results showed that the proposed method approach decreased the prediction error by 10%, and was able to choose the best or worst response with 40% accuracy, out of average of 14 responses.

1. はじめに

特定のタスクを持たず，雑談を行う非タスク遂行型の対話システムでは，対話用例を用いた用例ベース対話システムの研究が盛んに行われている [1]．用例ベース対話システムにおいて，用例は対話システムの品質を決定する重要な要素である．そのため，用例の収集と評価では用例ベースの品質を向上させるために様々な手法が提案されている．用例の収集では，人間同士の対話ログやドラマ・映画等の書き起こしスクリプト，SNS 等から得られた対話ログを用いて大規模な用例データベースを構築する手法が提案されている [2, 3, 4]．また，用例の評価においては，実際に用例を用いて行われた対話を分析することで，対話終了後に事後評価を行う手法が提案されている [5, 6, 7]．しかしながら，これらの手法を用いた場合，新たに用例を評価したい場合はその都度，人手による評価および人間と対話システムとの対話結果が必要であった．

本研究では，評価がアノテーションされた用例（既知用例）を学習データとして，評価がアノテーションされていない用例（未知用例）に期待される評価を推定する手法を提案する．本手法を用いることによって，対話システム構築以前に著しく評価を低下させることが予測される用例を除外することや，対話システム中で動的に本手法を利用することによって未知用例に対してもユーザの評価を最大化する応答を選択することが可能になる．

本稿では，実験的評価において，既知用例に対する交差検証を用いて予測値とアノテーション値との誤差を測ることで推定の精度を評価する．また，応答選択への有効性を検証するため，ある入力に対して複数の応答が期待できる用例のうち，評価値のアノテーションが最大の用例を提案法を用いて推定することが可能か検証する．

連絡先: 水上 雅博，奈良先端科学技術大学院大学，奈良県生駒市高山町 8916-5，0743-72-5265，masahiro-mi@is.naist.jp

2. 関連研究

本研究と関連した研究として対話中のユーザ満足度推定が挙げられる．対話中のユーザ満足度推定では，対話中のユーザの対話システムに対する反応から，対話システムへの満足度を推定する．Schmitt らはサポートベクターマシン (Support Vector Machine; SVM) を推定モデルに用い，音声認識の結果や認識結果の信頼度，音声から推定されたユーザの感情タグ，対話行為タグ，対話ターン数等を入力素性として，5 段階の対話満足度（Interaction Quality）を推定するモデルを提案している [5]．また，Higashinaka ら，Engelbrecht らは，対話におけるユーザの状態の移り変わりを隠れマルコフモデル (Hidden Markov Model; HMM) を用いてモデル化することで，満足度のみならず，スムーズさ，親密性，積極性などの推定も提案している [6, 7]．

これらの研究は，主な目的として既に終了した対話に対して，対話システムの事後評価となる満足度を推定するために用いられてきた．終了した対話を対象とすることで，ユーザの反応や音声，マルチモーダル情報などの多数の素性を利用することができる反面，実際に対話を遂行しなければならず，対話以前に期待できる評価を推定することもできなかった．

これに対して本研究では，対話システムが行った発話に対して予測されるユーザの評価を推定するため，対話システムを構築する際の用例の選別や，応答選択における評価最大化に利用することが可能である．その反面，対話中のユーザから得られる情報を利用することができないため，先行研究で有効とされた音声から推定されたユーザの感情タグ等の情報は評価値推定の素性として利用できない．

3. 対話用例の収集と評価値のアノテーション

3.1 シナリオベース対話用例の収集

本研究では，対話用例に対する評価の推定モデルの構築を行うために，文献 [8] で提案された対話用例収集を参考として，新たに複数名の被験者から用例を収集，快適度のアノテーショ

表 2: 用例収集の詳細

発話の収集	
被験者	7 名
収集発話数	41 発話 （重複は除く）
応答の収集	
被験者	15 名
収集応答数	511 発話
収集用例数	511 組

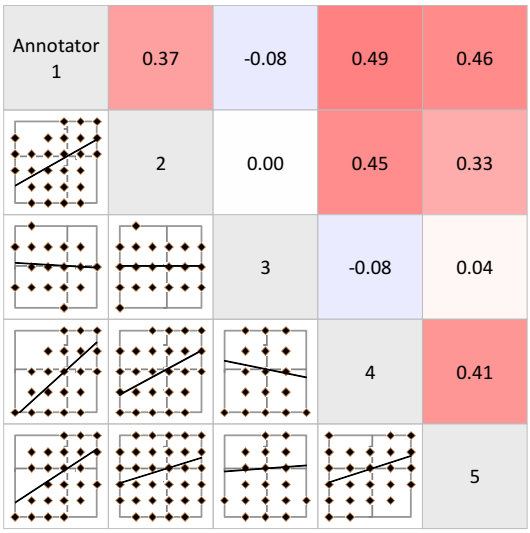

図 1: アノテータ間の相関

ンを行った．

まず，用例の収集では，一般的な社会人を対象とした自宅での生活シーン（例，帰宅時，夕食時，就寝時など）を計 14 シーン想定し，それに対して 7 人の被験者が各シーンに合致する発話を書き出し，計 42 発話を収集した．さらに，この 42 発話に対して，対話システムは何と応答すれば良いかを 15 人の被験者から収集し，入力と出力をペアとして用例を収集した．表 1 に収集された用例の一部を，表 2 に詳細を示す．

3.2 用例のアノテーション

3.1 節で収集された 511 組の用例に対して，5 名のアノテータが評価値として快適度を不快:1〜快適:6 の 6 段階でアノテーションを行った．なお，アノテーションは 1〜6 の 6 段階で行ったが，これ以降の分析，推定モデルの構築においては，アノテータ間のずれをできる限り少なくするため，アノテータごとのアノテーション値に対して Z スコア *1 への正規化を行ったものを用いる．用例にアノテーションされた快適度の分析としてアノテータ間の相関係数を散布図行列として図 1 に示す．

図 1 から，アノテータ 3 以外のアノテータ間には正の相関がみられるものの，全アノテータ間の平均決定係数 R^2 は 0.11 と低い値となった．このことから，用例に対する評価はユーザ依存性の高いスコアであるものの，アノテータによっては類似の傾向があると言える．

*1 Z スコアは集合の平均が 0，分散が 1 となる正規化スコア

4. 用例の評価値推定

2. 節で述べた関連研究である対話中のユーザ満足度推定では，対話中にユーザから得られた音声特徴量等の要素を用いて，ユーザの満足度を推定する．

これに対して本研究ではユーザから得られた情報を用いず，用例そのものからユーザが感じるであろう快適度を推定する．そのため，推定に用いることができる素性は限定される．

本研究では，ある用例 $\langle q, r\rangle$ にアノテーションされた快適度 $S_{\langle q,r\rangle}$ を推定する．目的変数としてアノテーションされた快適度 $S_{\langle q,r\rangle}$ を用い，説明変数に以下の素性を用いる．

- 用例の入力文 q の n-gram 頻度ベクトル
- 用例の出力文 r の n-gram 頻度ベクトル
- 用例の入力文 r の単語クラス頻度ベクトル *2
- 用例の出力文 r の単語クラス頻度ベクトル *2
- 用例の入出力文 q, r の単語共起頻度ベクトル
- 用例の入力文 q の中で単語感情極性対応表 [10] に存在する単語極性値の平均と最大最小値と存在しない場合のフラグ
- 用例の出力文 r の中で単語感情極性対応表 [10] に存在する単語極性値の平均と最大最小値と存在しない場合のフラグ
- 用例 $\langle q, r\rangle$ に快適度を付与したアノテータが誰かを示すフラグ

また，回帰モデルは文献 [5] における対話中のユーザ満足度推定を参考として，サポートベクター回帰 (Support Vector Regression; SVR)[11] による回帰モデルを学習した *3．学習された回帰モデルを用いて，未知の用例に対して期待される快適度を推定する．

5. 実験的評価

提案法の評価を行うために，既知の用例を 10 組に分割し，1 組ごとに一個抜き交差検証を用いて 4. 節の快適度推定を行った．評価の基準には，提案法によって推定された快適度と実際にアノテーションされている快適度の二乗誤差を求めた．本評価のベースラインとして，学習データにおける快適度の最頻値と実際にアノテーションされている快適度の二乗誤差を示す．なお，それぞれの結果の信頼区間を $p < 0.05$ の Bootstrap Resampling を用いて求める [12]．

図 2 から，アノテータごとに減少量に差はあるものの，ベースラインに比べて提案法では誤差は減少し，全体では誤差を 10%減少させることができた．提案法における誤差が最も大きかったのはアノテータ 3 を対象にした場合であった．図 1 で示した通り，アノテータ 3 は他のアノテータとは異なるアノテーション傾向を持っている．そのため，快適度推定ではこの傾向の差を学習できず，精度が向上しなかったと考えられる．

*2 当該の単語に対して，日本語 WordNet[9] から単語の持つ Synset ID を取得し，クラスとして与えた．

*3 SVR は学習データを 5 分割し，一個抜き交差検証によって最も誤差が小さくなるパラメータ C, γ と次元数を求めた．また，カーネル関数には線形カーネルと RBF カーネルを用いた．

表 1: 収集された用例の一部

入力（ユーザ発話）	出力（システム応答）
今日は何かあったっけ？	カレンダーを確認してみてください
今日は何食べようかな	寒いし，おでんなんかどうですか？

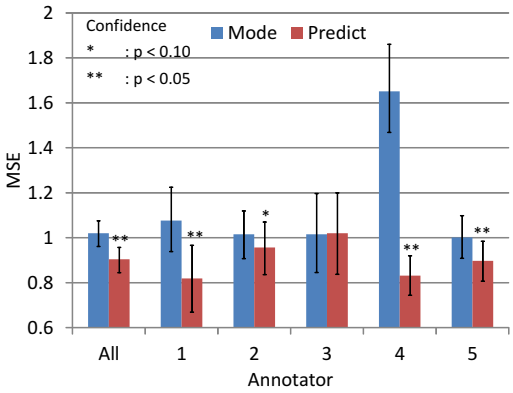

図 2: 快適度推定の精度

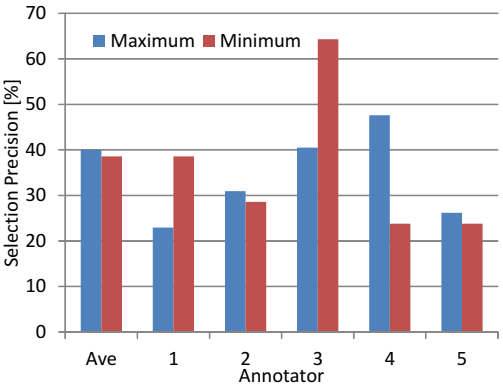

図 3: 応答選択の精度

次に，提案法における応答選択の有効性を検証するため，ある発話に対して複数の応答が該当する用例に対して，快適度が最大，最小のものに対して提案法が最大，最小の推定値を与えられたかどうか，すなわち，快適度最高，快適度最低の応答を選択することができたか評価する．この快適度最高，快適度最低の応答を選択する実験では，未知用例に対して快適度最高の用例を選ぶことで対話中のユーザに対する快適度を最大化したり，未知用例の中から快適度最低の用例を事前に除外することで，著しくユーザの快適度を損ねる用例を排除することを目的としている．

図 3 から，こちらもアノテータごとに差はあるものの，およそ 40%の精度で平均 14 個の用例の中から快適度最高，快適度最低の用例を選択することができた．快適度最高の用例を選択する精度が最も高かったのはアノテータ 4 を対象とした場合で，快適度最低の用例を選択する精度が最も高かったのはアノテータ 3 を対象とした場合であった．これに対して，快適度最高の用例を選択する精度が低かったケースはアノテータ 1，快適度最低の用例を選択する精度が低かったケースは 4 を対象とした場合であった．同一のアノテータの快適度最高，快適度最低での選択精度に差が出た原因として，本評価における選択精度に Precision を用いたため，快適度最高，快適度最低の用例が複数ある場合の評価がうまく行われなかったと考えられる．

6. まとめ

本研究では，対話システムに用いる用例に対して，既知用例から未知用例がユーザに与える快適度を推定する手法を提案し，実験を通してその性能を示した．

快適度の推定では，全体平均で 10%程度の誤差を減少させたが，アノテータによっては減少量が大きく異なることが分かった．この原因として，アノテータ間におけるアノテーションの傾向が異なる場合，他アノテータの評価データを学習に利用出来ず，回帰モデルの精度が低下するためと考えられる．また，アノテータに関する情報も単純なアノテータの識別番号のみであり，アノテータの特徴を示す情報が存在しなかったことが原因であると考えられる．

応答選択における有効性では，最高・最低の評価値を持つ用例をおよそ 40%の精度で選択することが可能であるが，こちらもアノテータによって選択精度が大きく異なることが分かった．また，今回は Precision による選択精度の評価を行ったが，最大・最小の評価値を持つ用例が複数存在する場合を考慮し，今後は Recall および F 値による評価が必要である．

対話システムにおいて，大規模な用例データベースを構築することと，ユーザ個々の好みや特徴を考慮して応答の快適度を推定，応答を選択することは非常に重要である．本研究で提案した快適度推定モデルに加え，対話中に推定されたユーザの情報を考慮することで，未知用例に対してもユーザの好みを考慮した応答文の快適度を推定することが可能であると考えられる．

参考文献

[1] C. Lee, S. Jung, J. Eun, M. Jeong, and G. G. Lee. A situation-based dialogue management using dialogue examples. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 1, pp. I–I. IEEE, 2006.

[2] L. Nio, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura. Developing non-goal dialog system based on examples of drama television. In *Natural Interaction with Robots, Knowbots and Smartphones*, pp. 355–361. Springer, 2014.

[3] E. Levin, R. Pieraccini, and W. Eckert. Using markov decision process for learning dialogue strategies. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, Vol. 1, pp. 201–204. IEEE, 1998.

[4] H. Murao, N. Kawaguchi, S. Matsubara, Y. Yamaguchi, and Y. Inagaki. Example-based spoken dialogue system using woz system log. In *SIGdial Workshop on Discourse and Dialogue*, pp. 140–148, 2003.

[5] A. Schmitt, B. Schatz, and W. Minker. Modeling and predicting quality in spoken human-computer interaction. In *Proc. SIGDIAL*, pp. 173–184, 2011.

[6] R. Higashinaka, Y. Minami, K. Dohsaka, and T. Meguro. Modeling user satisfaction transitions in dialogues from overall ratings. In *Proc. SIGDIAL*, pp. 18–27, 2010.

[7] K.-P. Engelbrech, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller. Modeling user satisfaction with hidden markov model. In *Proc. SIGDIAL*, pp. 170–177, 2009.

[8] 水上, 木村, 野村, G. Neubig, S. Sakti, 戸田, 中村. 対話システムにおける応答選択法の検討. 日本音響学会 2014 年秋季研究発表会 (ASJ), 北海道, 9 2014.

[9] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Enhancing the Japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 1–8, 2009.

[10] H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *Proc. ACL*, pp. 133–140, 2005.

[11] D. Basak, S. Pal, and D. C. Patranabis. Support vector regression. *Neural Information Processing-Letters and Reviews*, Vol. 11, No. 10, pp. 203–224, 2007.

[12] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pp. 388–395, 2004.