Vol.2018-NL-235 No.6
Vol.2018-SLP-121 No.6
2018/5/13

雑談対話システムのための未知語に頑健な興味推定手法

田中昂志, 高山隼矢, 荒瀬由紀,
大阪大学大学院情報科学研究科

{tanaka.koji, takayama.junya, arase}@ist.osaka-u.ac.jp

1 はじめに

近年，人間との雑談を目的とした対話システムが広く普及している．雑談対話システムが多くのユーザに長期的に利用されるためには，人間同士の対話と同様に，円滑な会話の実現が必要であり，そのためには心理的な働きとして社会的スキル [1] が求められる．社会的スキルの一つとして，各ユーザに対して出身地や性格等のユーザ情報に沿った会話を行うことが挙げられる．そのようなユーザ情報の中でも，特にユーザが興味のある話題は有効であると考えられる．雑談対話システムがユーザが興味のある話題を提供するためには，ユーザの発話中から興味のある語を特定する必要があり，既存研究 [2] では発話中の名詞に対するユーザの興味推定に取り組んでいる．

一方，対話システムの構築には深層学習が用いられることが多い [3] が，システムが保持できる語彙サイズには制約がある．ユーザが興味を持つような語は，流行を反映する新しい語や，一般語でないマニアックな語であることが多く，対話システムの保持する語彙に入っていないものが多いと考えられる．実際に我々が構築したアノテーションデータでは，興味ありと判断された語の 17.2% が Wikipedia コーパスにおいて頻度 100 以下の低頻度語であった．対話システムにおけるユーザの興味対象推定では，このようなシステムの語彙に存在しない未知語についてもユーザが興味を持っているかどうかを正しく判定できることが望ましい．しかし既存研究では，未知語については考慮されていなかった．

そこで，本研究では対話文における未知語に対してユーザが興味を持っているかを判定する手法を提案する．既存研究同様，興味推定の対象は名詞とし，以降では名詞かつ未知語であるものを「未知語」，それ以外のものを「既知語」と呼ぶ．興味を持つ語が発話に現れる場合,「アクション映画もサスペンス映画も面白かった.」のように興味対象を列挙するケース,「邦画なんかより洋画が良い.」のように興味対象とそうでない語を対比するケースが多いと考えられる．そこで，深層学習と Conditional Random Fields (CRF) [4] を組み合わせた系列ラベリングにより，発話テキストに現れる語に興味があるかどうかを周辺文脈を考慮しながら推定する．さらに，未知語に対して頑健な興味推定を実現するため，2 つの手法を用いる．未知語はそのベクトル表現がモデル中に存在しないことが問題となる．そこで未知語のベクトル表現を文字単位の Bidirectional LSTM (BiLSTM) [5] により生成する Character Embedding[6] を適用する．さらに，発話テキストに現れる名詞をタグに置き換え学習することで，興味対象について言及する際の表現の特徴を捉える．上記の「邦画なんかより洋画が良い.」の例では,「邦画」「洋画」の語が未知語であっても，周囲の「なんかより」「が良い」という表現から「洋画」にユーザが興味を持っていると推定できると期待できる．

対話文におけるユーザの興味をアノテートした既存データは存在しない．そこで本研究では，Twitter から収集した対話文 7,433 件について，クラウドソーシングにより発話テキストに現れる名詞にユーザが興味を持っているかどうかをアノテーションした．構築したアノテーションデータを用いて評価実験を行ったところ，提案手法は未知語に対する興味推定の F 値 86.6% を達成し，既存手法と比べて 9.1% 改善することが示された．

2 関連研究

本章ではユーザ発話における興味推定に取り組んでいる既存研究について述べる．Wang ら [7] は音声的特徴と言語的特徴を用いて話者がどの単語に興味があるかを推定している．言語的特徴は TF-IDF を用いており，未知語についても興味推定を行うことができる．この研究は与えられたスピーチ全体を入力として興味推定を行っており，十分な長さの文に対して TF-IDF を計算できる．しかし本研究で対象とするような,

Vol.2018-NL-235 No.6
Vol.2018-SLP-121 No.6
2018/5/13

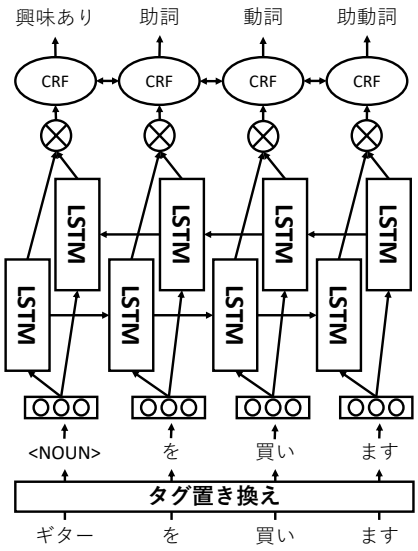

図 1: 提案手法の全体図

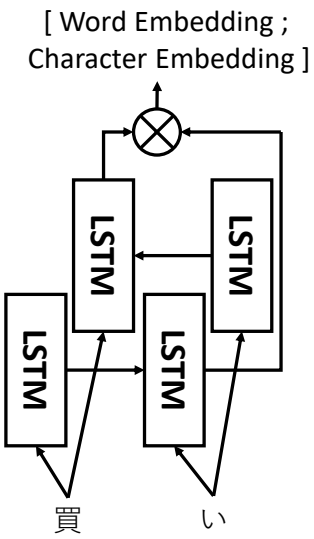

図 2: Character Embedding の概要

ユーザの 1 発話を入力とする場合，TF-IDF の値の信頼性を担保できない．

縄手ら [2] はユーザ発話外の名詞に対して興味推定を目的とした手法を提案しており，その過程において発話中における名詞の興味推定に取り組んでいる．ユーザ発話中の名詞とその周辺単語を文脈として抽出し，ニューラルネットワークに入力することで興味あり・なしの分類を行う．本研究と異なり，縄手らの手法では各名詞に対する興味推定は独立に行っている．また未知語の影響については考慮されていない．

3 提案手法

3.1 概要

提案手法の全体図を図 1 に示す．図 1 では，⊗ はベクトルの結合操作を表す．提案手法では，まず話者の発話中の名詞をタグに置換した後，入力の単語を Word Embedding と Character Embedding に変換し結合させたものをその単語の分散表現とする．そして BiLSTM-CRF モデル [6] に入力しする．正解データとして，名詞以外の品詞，すなわち動詞，形容詞，副詞，助詞，助動詞，接続詞，連体詞，感動詞，記号，接頭詞，フィラーに対してはそれぞれの品詞ラベルを用い，名詞については「興味あり」「興味なし」どちらかのラベルを用いる．

3.2 BiLSTM-CRF モデル

提案手法は系列ラベリング問題でよく用いられる BiLSTM と CRF で構成されている．図 1 のように，発話中の単語を逐次的に順方向の LSTM と逆方向の LSTM に入力し，各時刻においてそれぞれの LSTM から得られた出力を結合させ，文脈ベクトルを得る．文脈ベクトルを CRF 層に入力し，推定された前後の単語の品詞ラベルもしくは興味ラベルを考慮しながら，推定対象の品詞ラベルもしくは興味ラベルを推定する．

3.3 Character Embedding

提案手法では未知語のベクトルは保持していないが，文字単位の Embedding を用いてベクトルを生成することで，未知語のベクトルを近似する．図 2 のように単語内の文字を逐次的に双方向 LSTM に入力し，順方向の LSTM と逆方向の LSTM の最終出力を結合させたものを Character Embedding とする．これを Word Embedding と連結し，BiLSTM-CRF モデルの入力とする．

3.4 名詞のタグ置き換え

例えば,「邦画なんかより洋画が良い.」のような文中にある名詞の興味推定を行う場合,「邦画」「洋画」の語が未知語であり，それらが本来持つ分散表現を利用

表 1: アノテーションデータ例

SENTENCE: 邦画なんかより洋画が良い.
CANDIDATES:
邦画: 興味なし
洋画: 興味あり

できなくとも，周辺の「なんかより」や「が良い」という表現から,「洋画」にユーザが興味を持っていると推定できると期待できる．したがって，ある語が名詞であるという情報とその出現するパターンを学習させることで，興味の推定は可能と本研究では仮定する．そこで，名詞をタグ「<NOUN>」に置き換え，学習を行うことで未知語に対しての興味推定の頑健性を向上させる．

4 評価実験

4.1 興味推定のアノテーション

発話における話者の興味をアノテーションした既存データは存在しないため，本研究ではクラウドソーシングによりアノテーションを実施した．アノテートする対象として，対話システムに対する入力と同様口語体で記述されるツイートを用いた．Twitter[*1]から 10 単語以上からなるツイート 7,433 件を収集し，アノテーションを行った．Lancers[*2]でアノテータ 3 名を募集し，ツイート中の全名詞に対してツイートをしたユーザがその名詞に対して興味があるか否かのアノテーションを付与する．名詞の抽出には MeCab[*3]を使用し，MeCab の辞書に 2017/05/22 時点の mecab-ipadic-neologd[*4]を追加して使用する．表 1 に実際のアノテーションデータ例を示す．

アノテータはツイートとツイート内の名詞が与えられ，その名詞に対してツイートしたユーザが「興味がある」か「興味がない，もしくは日本語として意味をなさない」の 2 種類のラベルを付与する．アノテーションの基準は，ツイートしたユーザに「この名詞について興味があるか」と質問したときの予想回答とする．アノテータ間のラベル一致率を示す Fleiss' kappa の値は 0.18 であったが，2 者間のラベル一致率を示す Cohen's kappa は各アノテータペアについて 0.13, 0.32，0.33 であった．そこで最終的な正解ラベルはアノテータ 3 名の多数決により決定した．

[*1] https://twitter.com/
[*2] https://www.lancers.jp/
[*3] http://taku910.github.io/mecab/
[*4] https://github.com/neologd/mecab-ipadic-neologd

表 2: 実験データ

発話数（件）		7433
訓練データ	未知語の数（語）	5256
	既知語の数（語）	38563
開発用データ	未知語の数（語）	1267
	既知語の数（語）	5681
テストデータ	未知語の数（語）	1227
	既知語の数（語）	5382

4.2 実験設定

アノテーションデータを訓練用データ 5,953 件，開発用データ 740 件，テストデータ 740 件に分割した．また，Embedding 層の初期状態は Word2Vec の学習により得たものを使用する．Word2Vec の学習には 2017/11/22 時点の Wikipedia データを使用し，Word2Vec のウインドウサイズは 5，最小出現頻度は 100，ベクトルの次元数は 200 とした．Wikipedia データ内の語彙に含まれていない名詞を未知語とする．使用したデータの情報を表 2 に示す．

LSTM の中間層は 200 次元，ドロップアウト率は 30% とした．バッチサイズは 16 とし，Embedding 層の重みの更新は行うものとした．エポック数は 30 とし，開発用データで F 値を評価して最も高い F 値を示したエポック数でのモデルを使用した．

4.3 比較手法

実験では提案手法の未知語に対する興味推定の性能，また既知語に対する興味推定への影響を調査するため，テストデータ内の未知語のみを推定対象とした場合，テストデータ内の既知語のみを推定対象とした場合で比較を行った．

ベースラインとして，縄手らの手法及び単純な BiLSTM-CRF を用い，提案手法と比較する．縄手らの手法においては，未知語は正規分布に基づいた乱数を用いて生成したベクトルを未知語の単語分散表現とする．また，提案手法における各コンポーネントの効果を検証するため，Character Embedding を行わないモデルと名詞のタグ置き換えを行わないモデルを用いて精度の比較を行った．

4.4 実験結果

精度の比較に Precision，Recall，F 値の 3 つの指標を用いた．パラメータの初期化による影響を排除する

表 3: 実験結果（興味対象：未知語のみ）

手法	Precision	Recall	F 値
縄手ら	67.9 ± 0.9	90.3 ± 1.5	77.5 ± 0.1
BiLSTM-CRF	75.4 ± 1.4	69.6 ± 3.9	72.3 ± 1.6
提案手法	83.4 ± 1.7	$\mathbf{90.1 \pm 2.0}$	$\mathbf{86.6 \pm 0.4}$
-Character Embedding	75.2 ± 3.4	52.9 ± 10.9	61.4 ± 7.2
-タグ置き換え	$\mathbf{85.2 \pm 1.6}$	83.0 ± 4.3	84.0 ± 1.7

表 4: 実験結果（興味対象：既知語のみ）

手法	Precision	Recall	F 値
縄手ら	69.8 ± 0.9	$\mathbf{88.9 \pm 0.6}$	78.2 ± 0.4
BiLSTM-CRF	75.4 ± 1.7	84.4 ± 1.5	$\mathbf{79.6 \pm 0.4}$
提案手法	73.2 ± 2.2	80.1 ± 3.6	76.4 ± 0.5
-Character Embedding	52.2 ± 1.9	56.4 ± 12.9	53.5 ± 5.6
-タグ置き換え	$\mathbf{76.5 \pm 1.6}$	81.6 ± 2.7	78.9 ± 0.5

ため，各手法を 5 回ずつ訓練・評価し，95%信頼区間を計算した値を結果とする．

未知語の興味推定結果　表 3 に興味推定対象が未知語の場合の Precision，Recall，F 値を示す．推定対象が未知語の場合，提案手法が最も高い F 値である 86.6% を達成した．タグ置き換えを用いない場合，F 値は 2.5% 低下した．このことから，タグ置き換えの有効性が分かる．未知語であっても，興味のある対象を言述するときの表現のパターンを学習することで，興味推定を行えることが示された．また，BiLSTM-CRF と Character Embedding を組み合わせたモデル（タグ置き換えを行わない提案手法）が BiLSTM-CRF を F 値において 11.7% 上回る結果となった．このことより，Character Embedding が未知語の興味推定に大きく貢献していることが分かる．これは，Word Embedding では分散表現に変換できない未知語の近似がある程度有効に働いたためであると考えられる．

既知後の興味推定結果　表 4 に興味推定対象が既知語の場合の Precision，Recall，F 値を示す．推定対象が既知語の場合，BiLSTM-CRF が 79.6% の最も高い F 値となった．これより，名詞のタグ置き換えは推定対象が未知語の場合には精度向上に貢献するが，推定対象が既知後の場合には精度が低下する傾向にあることが分かる．これは，推定対象に「今日」や「の」などの明らかに興味の対象とならない一般語が含まれており，タグ置き換えによってこのような一般語を区別できなくなってしまうためと考えられる．また学習した単語ベクトルが存在する既知語については，Character Embedding は貢献しないことが分かる．以上の結果から，既知語については BiLSTM-CRF，未知語については提案手法を用いることで，既知後における推定性能を維持したまま，未知語に対する推定性能を向上できると考えられる．

表 5 は提案手法と 縄手らの手法を用いて興味推定を行った例である．表 5 より，提案手法では未知語である「www」,「カミコベ」を正しく興味推定できていることが分かる．一方で，縄手らの手法では「www」は正しく興味推定できているが,「カミコベ」は誤った推定をしている．これは，縄手らの手法では文脈中に存在する未知語（「www」）の情報を用いることができず，興味推定が困難であったためと考えられる．

5　まとめ

本研究では，ユーザ発話中の未知語に対する頑健な興味推定を目的とし，BiLSTM-CRF にタグ置き換え，Character Embedding を適用した手法を提案した．

ツイート中の名詞について興味あり・なしのラベルをクラウドソーシングにより付与した 7,433 件のデータを用いて行った評価実験の結果，提案手法により未知語の興味推定において 86.6% の F 値を達成できることを示した．

今後の課題として，Character Embedding をサブワードに拡張したモデルの検討，及び興味推定に基づいた応答を生成する対話システムの構築に取り組む予定である．

6　謝辞

本研究は，Microsoft Research Asia 及び株式会社コトバデザインの助成を受けたものです．研究において有益な助言を頂いた大阪大学データビリティフロンティア機構の Chenhui Chu 特任助教，大阪大学大学院情報科学研究科の五十川真生氏，野本英梨子氏に感謝します．

参考文献

[1] 大坊郁夫．コミュニーケーション・スキルの重要性．**日本労働研究雑誌**, 48(1):13–22, Jan. 2006.

[2] 縄手 優矢，稲葉 通将，高谷 智哉，高橋 整，山田 健一．雑談対話ログを用いた話者の潜在的興味対象の推定．**人工知能学会全国大会**, pages 8–12, May 2017.

表 5: 「心配させんなよ wwwってかカミコベであえなかったじゃねぇか www」と入力した場合の興味推定結果

推定対象	提案手法	縄手ら	正解ラベル
心配	興味なし	興味なし	興味あり
www	興味なし	興味なし	興味なし
カミコベ	興味あり	興味なし	興味あり
www	興味なし	興味なし	興味なし

[3] Oriol Vinyals and Quoc V. Le. A neural conversational model. *In Proc. of the ICML Deep Learning Workshop 2015*, abs/1506.05869, July 2015.

[4] John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proc. of the International Conference on Machine Learning (ICML)*, 8:282–289, 6 2001.

[5] Alex Graves and Jrgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602 – 610, Aug. 2005. In Proc. of International Joint Conference on Neural Networks (IJCNN).

[6] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proc. of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 260–270, June 2016.

[7] William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Computer Speech and Language*, 27(1):168–189, 2013.