人の言語使用における単語の意味の揺らぎの解明に向けて

大葉 大輔 * 佐藤 翔悦 * 赤崎 智 * 吉永 直樹 † 豊田 正史 †
* 東京大学大学院 情報理工学系研究科 † 東京大学 生産技術研究所
{oba, shoetsu, akasaki, ynaga, toyoda}@tkl.iis.u-tokyo.ac.jp

1 はじめに

人は様々な感覚機能で感じた世界を言語によって表現するが，個人が言語で表現する意味には本質的に揺らぎが存在する．このような個人の言語使用における単語の意味の揺らぎは，日常生活において他者への適切な情報伝達を阻害するだけでなく，計算機で様々な自然言語処理タスクを解く際にも問題を引き起こす．例えば，ある製品の評価文書から対象の製品を推定するような，**客観的な出力**を求めるタスクでは，文書中の単語の意味が推定結果に影響を与えるため，個人ごとの単語の意味の揺らぎを考慮した上でタスクを解くことが望ましい．

本稿では人の言語使用における単語の意味の揺らぎ ((personal) semantic variation) を分析することを目標に，個人の評価文書から評価対象を推定するタスクを通じて，意味の揺らぎを personalized word embedding で捉える手法を提案する．具体的には，**(1) 評価者ごとのパラメタの fine-tuning**，及び **(2) 評価対象の様々な属性を推定するマルチタスク学習**を行う（図 1）．

実験では，ratebeer.com から収集したビールの評価文書集合を用いて，提案手法が評価対象の推定に効果的であることを確認する．さらに，獲得した personalized word embedding を分析することでどのような単語が強い semantic variation を持つか詳細に分析する．

2 関連研究

これまで，主に感情分析 [1] や対話生成 [2]，機械翻訳 [3] などのタスクを対象に，タスクの入出力に関与する人物を考慮してモデルの推定性能を向上する研究が行われている．しかし，これらの研究はいずれも**主観的な入力から主観的な出力**を推定する設定（例: 評価文書から評価対象の評価極性を推定，また対話で発話からの応答を推定）のもと行われており，結果としてモデルは入力を生成した個人の言語使用における意味の揺らぎ (semantic variation) のみならず，出力ラベルの**注釈バイアス**（注釈者ごとの出力ラベルの付け方の偏り）や**選択バイアス**（評判分析における評価対象の恣意的な選択に起因する，出力ラベルの偏り）[1] も考慮することとなる．結果として，得られたモデルをもとに本研究で対象とするような semantic variation のみを分析することは困難である．

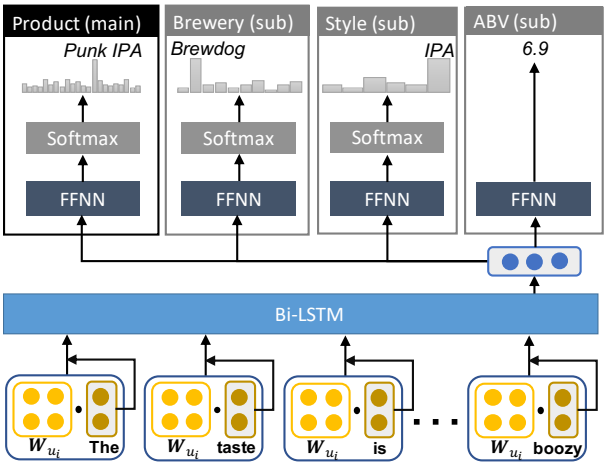

図 1: 評価対象の推定に用いる深層学習モデル．

本研究で扱う評価文書の評価対象を推定するタスクには注釈バイアスは存在しない．また選択バイアスについても，同一評価者が同一対象を一度のみ評価するデータセットを用いることで回避できる．

3 提案手法

本節では，単語の semantic variation を考慮した personalized word embedding を獲得するための深層学習モデル（図 1）について説明する．基本的なアイデアとして，評価者ごとに独立した単語埋め込み層を持つ深層学習モデルで評価文書から評価対象を推定する問題を解き，personalized word embedding を学習する．

評価対象の推定に用いる深層学習モデルを図 1 に

示す．本モデルは，入力の評価文書の各単語を単語ベクトルへと変換し，それらを Bi-directional Long-short Term Memory (Bi-LSTM) で潜在表現に変換して Feed-forward 層と Softmax 層から構成される後続ネットワークへと入力し，評価対象の推定を行う．

提案モデルの工夫として，一度の学習で膨大な評価者に対するパラメタを同時に学習することは空間効率的に現実的でないため，まず全評価者を同一視してモデルを事前学習した後，評価者固有のパラメタのみを各評価者ごとに fine-tuning するというアプローチを取る．さらに，評価対象の推定という超多クラス分類問題の学習をなるべく安定して行うために，評価対象の属性推定を補助タスクとしたマルチタスク学習を行う．それぞれ，以下で順に説明する．

3.1 評価者ごとのパラメタの fine-tuning

まず, モデルパラメタの事前学習を行う．提案モデルは入力文書中の単語 w_i を d 次元のベクトル $\boldsymbol{e}_{w_i} \in \mathbb{R}^d$ に変換する．事前学習では，この $\boldsymbol{e}_{w_i}$ を評価者共通の変換行列 $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ とバイアスベクトル $\boldsymbol{b} \in \mathbb{R}^d$ を用いて変換し，活性化関数 ReLU を適用した後，$\boldsymbol{e}_{w_i}$ と足し合わせて後段の層に入力する．

$$\text{ReLU}(\boldsymbol{W} \cdot \boldsymbol{e}_{w_i} + \boldsymbol{b}) + \boldsymbol{e}_{w_i} \tag{1}$$

$\boldsymbol{e}_{w_i}, \boldsymbol{W}, \boldsymbol{b}$ は後段の層のパラメタと共に，推定タスクの損失を用いて最適化する（事前学習）．

次に評価者固有のパラメタを学習するために，各評価者 u_j が書いた評価文書のみを用いて $\boldsymbol{W}$ と $\boldsymbol{b}$ の fine-tuning を行う．評価者ごとの personalized word embedding $\boldsymbol{e}^{u_j}_{w_i}$ は $\boldsymbol{W}$ と $\boldsymbol{b}$ を $\boldsymbol{W}_{u_j}, \boldsymbol{b}_{u_j}$ と再定義した形で以下のように計算する．[1]

$$\boldsymbol{e}^{u_j}_{w_j} = \text{ReLU}(\boldsymbol{W}_{u_j} \cdot \boldsymbol{e}_{w_i} + \boldsymbol{b}_{u_j}) + \boldsymbol{e}_{w_j} \tag{2}$$

学習では，$\boldsymbol{W}_{u_j}$ と $\boldsymbol{b}_{u_j}$ 以外のパラメタを固定して fine-tuning することで，semantic variation を $\boldsymbol{e}^{u_j}_{w_i}$ に押し込めることを狙う．なお，他のパラメタが固定されているため，$\boldsymbol{e}^{u_j}_{w_i}$ は異なる評価者間で比較可能である．

また，評価者固有のパラメタである $\boldsymbol{W}_{u_j} \in \mathbb{R}^{d\times d}$ と $\boldsymbol{b}_{u_j} \in \mathbb{R}^d$ は各評価者ごとに独立して（空間効率良く）学習することが可能である．そのため，本手法は大規模な評価者集合に対しても容易にスケールする．

[1] personalized word embedding の獲得のための非線形変換では個人の言語感覚の違いを捉えて元の単語からの意味の差分を学習することが目的であるため，式 (2) のように residual ネットワークを用いて $\boldsymbol{e}_{w_i}$ からの差分を表現する構造にした．

3.2 マルチタスク学習による学習の安定化

本研究で解くタスクは出力ラベル空間が広大であるため，限られた訓練データでは学習が安定しないことが予想される．そこで我々は，評価対象を抽象化した評価対象の属性（例:対象のカテゴリ，製造元）の推定が評価対象の推定に寄与すると考え，各属性の推定問題を補助タスクとしてマルチタスク学習により解くことで，主タスクである評価対象の推定の性能向上を図る．具体的には，各補助タスクごとに図 1 のように Feed-forward 層と Softmax 層から構成されるネットワークを追加し，全タスクの出力に対する損失を単純に足し合わせたものを最適化する．なお，選択バイアスの混入を避けるため，マルチタスク学習はモデル全体の事前学習時のみ行い，fine-tuning 時は主タスクのみを考慮して評価者固有のパラメタの最適化を行う．

4 実験

本節では，提案手法を実評価文書に対して適用する．semantic variation を捉えた評価対象推定モデルは, 評価対象をより良く推定できると考えられるため, モデルの推定精度について確認する．更に獲得した personalized word embedding について詳細に分析し，単語に現れる semantic variation の程度や傾向を分析する．

4.1 実験設定

データセット　我々は，ratebeer.com[2]から収集されたビールの評価文書 [4] を評価対象の推定タスクのデータセットとして用いる．データセットはビールについての総数 2,924,163 件の評価文書，評価対象として 110,369 種類のビールが含まれる．3.2 節のマルチタスク学習で製品と同時に推定する属性としては，各ビールに対して付与されている**スタイル (style)**，**醸造所 (brewery)**，および**アルコール度数 (ABV)** を用いた．本研究では評価者ごとに十分な量の学習データを担保するため，評価を 100 件以上書いている評価者 3670 人による 109,912 種類のビールに対する 2,695,615 件の評価文書を抽出し，学習 (2,156,493)，開発 (269,561)，評価 (269,561) データとして 8:1:1 の割合で分割し，学習および評価に用いた．なお，このデータセットには 89 種類のスタイルと 7,507 種類の醸造所が含まれていた．

[2] https://www.ratebeer.com

Bi-LSTM		最適化	
層数	1	バッチサイズ	200
隠れ層の次元数	200	ドロップアウト率	0.2
単語埋め込み層の次元数	200	最適化手法	Adam
語彙サイズ	100,288	(初期学習率)	0.0005

表 1: 評価対象推定モデルの主なハイパーパラメタ.

モデル		主タスク	補助タスク		
multi-task	personalize	product [Acc.(%)]	brewery [Acc.(%)]	style [Acc.(%)]	ABV(%) [RMSE]
		15.74	n/a	n/a	n/a
	✓	16.69	n/a	n/a	n/a
✓		16.16	(19.98)	(49.00)	(1.428)
✓	✓	**17.56**	(**20.81**)	(**49.78**)	(**1.406**)
ベースライン		0.08	1.51	6.19	2.321

表 2: 評価文書の評価対象の推定タスクの実験結果. ベースラインは，分類問題で最頻クラス，回帰問題で平均値を出力.

モデルとハイパーパラメタの設定　3.1 節で述べた評価者固有のパラメタの fine-tuning，および 3.2 節で述べたマルチタスク学習について，それぞれ有効/無効化した 4 種の深層学習モデルを PyTorch (ver. 0.4.0)[3]で実装し，前述の学習データを用いて学習した．表 1 にモデルの学習に用いたハイパーパラメタを示す．モデルの入力 e_i は，4.1 節で述べたデータセットで Skip-gram [5] を学習して得た単語ベクトルを用いて初期化する．この際の語彙集合は元のデータセットにおいて 10 回以上使用された 100,288 語に設定した．

全評価者を同一視してモデル全体のパラメタをマルチタスク学習する際の損失は，分類タスクにクロスエントロピー損失，回帰タスクに二乗損失を用い，これらを足し合わせて用いた．パラメタの最適化には Adam を用い，最大 100 エポックまで最適化を行い開発データで最も高い性能となるエポックのモデルを評価に用いた．評価者ごとのパラメタは，この性能最大のエポックのモデルで評価者非依存のパラメタを固定し，主タスクの損失のもと，評価者ごとに $\boldsymbol{W}_{u_j}, \boldsymbol{b}_{u_j}$ を fine-tuning した (式 (2)).

4.2　実験結果

表 2 に評価データにおける分類タスク (product, style, brewery) の正解率 (Accuracy) と回帰タスク (ABV) の平均二乗誤差 (RMSE) を示す．主タスクにおいては評価者ごとのパラメタの fine-tuning およびマルチタスク学習はそれぞれ有効であり，その効果は相乗的であった．主タスクに対する評価者パラメタの fine-tuning により補助タスクの性能も向上していることは注目に値する．以上の結果より, semantic variation をより良く考慮できていると考えられる主タスクで最良の結果を示したモデルで獲得された personalized word embedding を以降の分析対象に使用する.

[3] https://pytorch.org/

top-50
ery **bready** ark **slight floral toasty tangy updated citrusy soft deep** mainly **grassy** aroma **doughy dissipating** grass ot **great earthy** smell **toasted** somewhat **roasty soapy** perfume **flowery lingering musty** citrus **malty** background malt present hue **minimal** earth **foamy** faint **dark medium clean nice** copper hay bread herbs **chewy** complexity toast reddish
bottom-50
reminds cask batch oil reminded beyond **canned conditioned** double abv hope horse oats rye brewery blueberry blueberries maple bells **old** cork shame dogfish become dog hand **plastic** course remind christmas cross rogue **extreme organic fat** lost words islands etc growler **hot** heat stout alcohol unibroue pass nitro **longer** scotch **rare**

表 3: Semantic variation の大きい（小さい）単語（太字は形容詞）.

4.3　personalized word embedding の分析

提案手法により獲得した personalized word embedding を分析することで，どのような単語が semantic variation を伴うかを確認する．対象とする単語は語彙中の全単語から機能語と記号類を除いたもののうち，3 割以上の評価者が少なくとも 1 度用いたものに限定する．単語 w_i の semantic variation（の大きさ）を以下の式で定義する.

$$\frac{1}{|\mathrm{U}(w_i)|} \sum_{u_j \in \mathrm{U}(w_i)} (1 - \cos(\boldsymbol{e}_{w_i}^{u_j}, \overline{\boldsymbol{e}}_{w_i})) \qquad (3)$$

ここで，$\boldsymbol{e}_{w_i}^{u_j}$ は評価者 u_j の単語 w_i に対する personalized word embedding で，$\overline{\boldsymbol{e}}_{w_i}$ は単語 w_i を用いた評価者の集合 $U(w_i)$ について e_w^u を平均したものである.

表 3 に semantic variation が大きい（小さい）単語の例を示す．semantic variation が大きい語の多くは事物の性質の程度を表現する形容詞であり，直感に即した結果となっている.

次に，単語の通時的な意味変化 [6] および地理的な意味変動 [7] の分析で用いられた，単語の頻度 (**frequency**), 使用者の割合 (**dissemination**), 語義数 (**polysemy**) という 3 つの観点で semantic variation を分析し

(a) 頻度 (frequency)　(b) 使用者の割合 (dissemination)　(c) 語義数 (polysemy)

図 2: 各単語の personalized word embedding の semantic variation．各指標との pearson の相関係数はそれぞれ (a) **0.43**, (b) **0.29**, (c) -0.07 である．トレンドラインはブートストラップカーネル回帰による 95%信頼区間を示す．

た．dissemination については対象とした 3670 評価者のうち少なくとも 1 度は対象の単語を用いた評価者の割合を用い，polysemy については単語の WordNet [8] における synset の数とした．

図 2 にそれぞれの尺度と，semantic variation との大きさの相関を示す．興味深い点として，[6, 7] で報告された結果とは異なり，semantic variation は frequency および dissemination と高い相関を示した．その一方で，polysemy との相関は低いものとなっている．この理由としては次のように考えられる．[6, 7] で用いられた一般のドメインのデータセットとは異なり，本研究で用いたデータセットはビールに関するドメインに属するため，“soft” や “soapy” といった形容詞が高頻度で用いられる．またこれら形容詞の使用は個人の言語直感に依存するため，frequency, dissemination が semantic variation との高い相関を示したと考えられる．その一方で本研究で用いたデータセットはドメインが限定されており，文書中の単語はあくまでビールを表現するために用いられることから，語義そのものが変わるような場合は少ない事が polysemy との相関の低さの理由であると考えられる．

5　おわりに

本稿では，個人の言語使用における単語の意味の揺らぎ (semantic variation) を解明するため，個人の評価文書から評価対象を推定するタスクを解くことで，semantic variation を考慮した personalized word embedding を獲得する手法を初めて提案した．semantic variation を適切に捉えるため，提案手法では評価者ごとのパラメタの fine-tuning および評価対象の属性を推定するマルチタスク学習を行う．ビールについての評価文書集合を用いた実験により，提案手法は評価対象の推定に有効であることを確認し，獲得した personalized word embedding からどのような単語にどのような傾向で semantic variation が現れているかを分析した．

今後は，多様なドメインのデータセットで semantic variation の傾向を検証する予定である．また，性別や年齢などの評価者の属性と semantic variation の関係に関する分析を行いたい．

謝辞　本研究は JSPS 科研費 16H02905 の助成を受けたものです．

参考文献

[1] Wenliang Gao, Naoki Yoshinaga, Nobuhiro Kaji, and Masaru Kitsuregawa. Modeling user leniency and product popularity for sentiment classification. In *IJCNLP2013*, pp. 1107–1111, 2013.

[2] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *ACL2016*, pp. 994–1003, 2016.

[3] Joern Wuebker, Patrick Simianer, and John DeNero. Compact personalized models for neural machine translation. In *EMNLP2018*, pp. 881–886, 2018.

[4] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *ACM2013*, pp. 165–172, 2013.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS2013*, pp. 3111–3119, 2013.

[6] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL2016*, pp. 1489–1501, 2016.

[7] Marco Del Tredici and Raquel Fernández. Semantic variation in online communities of practice. In *IWCS2017*, 2017.

[8] George A Miller. Wordnet: a lexical database for english. *ACM1995*, Vol. 38, No. 11, pp. 39–41, 1995.