1L-07

単語の概念関係を考慮したレビューの評価観点自動抽出手法

小森 雄太 † 木村 優介 ‡ 楠 和馬 ‡ 波多野 賢治 †
† 同志社大学文化情報学部 ‡ 同志社大学大学院文化情報学研究科

1 はじめに

レビューは，実際の体験から得たユーザの率直な意見を含んでいるため，閲覧時の有用な情報になり得るとして注目されている．しかし，膨大なレビューの閲覧負担が問題視されており，閲覧負担の軽減に有効な手段の一つである，評価観点に基づいたレビューの提示が求められている．評価観点とは表 1 に示した例のように，評価が述べられている観点を指し，評価の対象を明示している具体的な単語を評価対象語という．このような背景から，閲覧負担軽減を目的とし，レビューから評価観点を抽出する研究が数多く行われている．

近年の評価観点抽出手法では，ルールベース・教師あり手法のデータ整備コストを減らすため，教師なし手法が提案されている [1]．この手法では，オートエンコーダにより評価観点の分散表現を学習し，学習した評価観点ベクトルの近傍に位置する単語を評価対象語として抽出することで，評価観点を表現する集合を作成する．次に，得られた集合から評価観点名を付与するために，集合の意味解釈を行う．しかし，評価対象語集合を人手で意味解釈し，評価観点となる名称を付与する必要があり，依然として人的コストを要する．

そこで本研究では，具体的な語の集合である評価対象語集合を，それらを包含する抽象的な一単語で表現することで，評価観点を機械的に付与することを目指す．

2 関連研究

文書分類を目的としたクラスタリング結果の意味解釈を機械的に行った研究として，上位概念語を用いた研究がある [2]．この手法では，作成した文書集合から複数の代表語を抽出し，各代表語に対する上位概念語をその集合のラベルとして複数付与する．上位概念語は代表語の意味を包含する抽象的な語であるため，類似語集合の意味を把握することができる．その結果，文書集合の意味解釈を簡易化できたと報告している．

しかし，閲覧負担軽減を目的とした評価観点名の付与を行うためには，評価観点に含まれる情報の把握を適切に行えるように，複数の上位概念語ではなく評価観点名を一語に同定することが望ましい．また，単に評価対象語集合を包含する上位概念語という条件のみ考慮し，過度に抽象的な評価観点を付与してしまった場合，ユーザがその評価観点に含まれる内容を判断できない．そのため，包含関係と理解の容易さを併せ持つ適当な上位概念を付与することは困難である．

表 1: 評価を含むレビューの例

レビュー例	評価観点	評価対象語
I eat a great pizza.	Food	pizza
The server was friendly.	Staff	server
The station is close.	Location	station

3 提案手法

本研究では，評価対象語集合の内容を表現する適切な評価観点を表す語を付与するために，集合内で単語の概念関係を考慮し，より上位に位置する語の抽出を提案する．関連研究では，上位概念語の付与を文書内では明示されていない語を用いて行っていた．しかし，レビューでは評価を記載する際に，具体的な評価対象語だけではなく，抽象的な評価観点も用いられる．そのため，多くの評価観点はレビュー内で明示されており，作成した評価対象語集合にも含まれている．

そこで，評価対象語集合からより上位概念に位置する語の抽出を行う．その結果，評価観点としてレビューで使用されることのない過度に抽象的な語を抽出する可能性を下げ，評価対象語内でより多くの意味を包含する一語を選択することができる．

評価対象語間の概念関係を把握するために，単語の概念関係を双曲空間で学習し，分散表現を作成することができる Poincaré Embeddings [3] を使用する．Poincaré Embeddings は，双曲空間の持つ性質を利用し，階層関係における単語間の距離を保ったまま，単語を分散表現化することができる．学習の結果得られる分散表現は，空間内で中心部は抽象度が高く，周縁部に位置するほどより具象度が高くなるように埋め込まれる．そのため，学習した各単語の分散表現と空間内の中心間距離を計測することでより上位に位置する語を抽出することができる．ここで，$\mathbb{R}^d$ を d 次元の実ベクトル空間，$\|\boldsymbol{x}\|$ をユークリッドノルムとすると，双曲空間内

An Extracting Method of Reviews' Aspects based on Lexical Conceptual Relations
†KOMORI Yuta, ‡KIMURA Yusuke, ‡KUSU Kazuma, †HATANO Kenji
†Faculty of Culture and Information Science, Doshisha University ‡Graduate School of Culture and Information Science, Doshisha University

のデータは d 次元の単位球 $\mathcal{B}^d = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| < 1\}$ 内で表現され，その空間における二点 $\boldsymbol{u}, \boldsymbol{v}$ の双曲線距離は逆双曲線関数 arcosh を用いて次式で与えられる．

$$d(\boldsymbol{u}, \boldsymbol{v}) = \operatorname{arcosh}\left(1 + 2\frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{(1 - \|\boldsymbol{u}\|^2)(1 - \|\boldsymbol{v}\|^2)}\right) \quad (1)$$

また Poincaré Embeddings は，学習データに上位・下位概念語の対を必要とする．そのため，本研究では概念辞書の WordNet[1] から学習データを作成する．各評価対象語を起点とし，WordNet 上でその上位概念，下位概念に位置するすべての語を抽出する．その際，WordNet に登録されていない語は具体的な下位概念語である可能性が高く，評価観点には適さないため無視する．そして，作成した上位・下位概念語の対を positive samples，同データ内で単語の出現頻度に比例する確率を基にランダム抽出した二単語を negative samples として学習を行う．このように学習した評価対象語間の概念関係から，評価対象語集合内でより上位に位置する一語の抽出に取り組む．

4 評価実験

提案手法の有効性を確認するため，人手で付けたラベルにどれだけ近いラベル付けができたかを三手法で比較し検証する．そこで，レストランのレビューを収集した Citysearch corpus [2] と He et al.[1] によって付与された評価観点を用いて評価実験を行う．He et al. が作成した評価観点の正解ラベルは，レビューから抽出した評価対象語集合に付与した 14 種類の観点があり，その内「その他」を意味する二つを除いた 12 種類の観点を使用する．

類似度の計算には，非文脈依存の分散表現作成方法で最も精度がよいとされる fastText [3] でレビューを学習し作成した分散表現を各評価観点に割り当て，単語間のコサイン類似度を求める．複合語，未知語の分散表現は N-gram の情報を活用した fastText のサブワードを用いて作成する．

比較対象として，各集合で全単語の一階層上の概念を WordNet で抽出し， TF-IDF で重み付けして求めた単語（WordNet），学習した評価観点ベクトルの最近傍（ベースライン）を用いる．これらと比較することで，評価対象語内から評価観点を抽出する有効性と上位概念語を用いる有効性を確かめる．

また，評価対象語集合を形成する際の最適な単語数が不明であるため，20 から 100 語の間で 10 語ずつ単語数を増加させて評価実験を行い，平均類似度が最も高くなった単語数を最適な評価対象語数とする．

[1]Princeton University,“About WordNet.”, https://wordnet.princeton.edu/, (2021/1/8 閲覧)

[2]Gayatree Ganu,“Restaurant Reviews Dataset”, http://www.cs.cmu.edu/~mehrbod/RR/, (2021/1/8 閲覧)

[3]Facebook,“fastText”, https://github.com/facebookresearch/fastText, (2021/1/8 閲覧)

表 2: 評価観点と正解ラベルの平均コサイン類似度

	提案手法	WordNet	ベースライン
平均類似度	**0.4434***	0.1278	0.2386
分散	0.0941	0.0063	0.0363

表 2 に実験の結果得られた各評価対象集合と正解ラベルの平均コサイン類似度を記す．表中の * は有意水準 5%で対応のある t 検定を行った結果 WordNet との有意差があったことを示す．提案手法は二つの比較対象よりも高い平均コサイン類似度となったが分散が大きく評価対象語集合により性能に散らばりがあった．各評価観点別に見た場合，コサイン類似度が低くなった評価観点は，評価対象語の中に含まれる同綴異義語が実際とは異なる概念関係として学習されたためではないかと考える．これは評価観点ベクトルの学習を行う過程で Word2Vec を用いて分散表現を作成していたため，同綴異義語が同じ分散表現を持つことに起因すると考えられる．評価対象語数を変化させて行った実験では，評価対象語数が 50 のときに最も類似度が高くなった．

5 おわりに

本研究では，レビューの閲覧負担軽減を目的とし，評価観点の抽出手法に関する提案を行った．評価実験の結果，提案手法により比較的人の感性に近い評価観点をレビューから自動的に抽出することができた．また，レビューにおいては明示された評価対象語から評価観点を作成できることが示唆された．

今後は，より尤もらしい評価観点を付与するために文脈依存の分散表現を考慮し同綴異義語を区別して学習する手法を提案し，WordNet に含まれず学習できなかった単語の影響を調査する必要がある．

参考文献

[1] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An Unsupervised Neural Attention Model for Aspect Extraction. In *Proc. ACL2017*, pp. 388–397. ACL, 2017.

[2] Yuen Hsien Tseng, Chi Jen Lin, Hsiu Han Chen, and Yu I Lin. Toward Generic Title Generation for Clustered Documents. In *Proc. AIRS2006*, pp. 145–157. Springer, 2006.

[3] Maximillian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems 30*, pp. 6338–6347. Curran Associates Inc., 2017.