視覚特徴と言語特徴を用いた単語の心像性推定の検討

松平 茅隼† カストナー マークアウレル† 井手 一郎† 川西 康友†
平山 高嗣† 道満 恵介†‡ 出口 大輔† 村瀬 洋†
†名古屋大学 ‡中京大学
{matsuhirac, kastnerm}@murase.is.i.nagoya-u.ac.jp
{ide, kawanishi, murase}@i.nagoya-u.ac.jp
{takatsugu.hirayama, ddeguchi}@nagoya-u.jp
kdoman@sist.chukyo-u.ac.jp

1 はじめに

心理言語学の分野において，Paivio らにより単語の「心像性」が定義されている [1]．心像性とは，単語のイメージしやすさを表す指標である．例えば,「犬」や「車」などイメージが湧きやすい単語の心像性は高く,「平和」や「知識」などイメージが湧きにくい単語の心像性は低くなる．このような心像性の応用例として，画像とそのキャプションの関係の定量的分析 [2] などが挙げられる．

一般に心像性は被験者実験を通じて定量化されるが，未知語を含む任意の単語に対応するためには，心像性を推定する手法が必要となる．そこで本研究では，単語に関する視覚的な特徴と言語的な特徴を手がかりにして心像性を推定する手法を提案する．ここでは，視覚的な特徴として画像から抽出する特徴を，言語的な特徴としてテキストデータから抽出するテキスト特徴及び単語の発音に着目した発音特徴を考える．

以降，2 節で関連研究を紹介し，3 節で心像性の自動推定手法を提案する．次に 4 節では，提案手法の評価実験を行う．最後に 5 節で本発表をむすび，今後の課題について述べる．

2 関連研究

Paivio らは，具象性（concreteness）や有意性（meaningfulness）と共に，単語の心像性（imageability）を定義した [1]．心像性とは，単語のイメージしやすさを表す指標である．

1 節で述べた問題意識に基づいて，Kastner らは SNS 上の画像データを用いて単語の心像性を推定する手法を提案した [3]．彼らは単語の心像性と SNS 上の画像データの関連性を仮定し，SNS 上の画像データのマイニングにより視覚特徴のみを用いて心像性を推定した．

一方，Ljubešić らは大規模なコーパスで事前学習をした fastText [4] による単語の分散表現を使用し，言語特徴のみを用いて単語の心像性及び具象性を推定した [5]．

3 単語の心像性推定手法

本節では，視覚特徴と言語特徴を用いた単語の心像性推定手法を提案する．図 1 に示すように，処理手順として，まず単語に関連する画像データ及びテキスト・発音データから視覚特徴と言語特徴をそれぞれ抽出し，次にそれらを統合し，最後に回帰モデルにより心像性を推定する．

3.1 視覚特徴の抽出

Kastner らの手法 [3] を踏襲して，視覚特徴の抽出には SNS 上の画像データを利用する．まず SNS 上の画像付き投稿データに付随するタグなどを手がかりにして，1 単語につき画像を 5,000 枚収集する．そして，各画像について画像特徴を抽出する．この際，低レベルの特徴量（HSV 色ヒストグラム，GIST 特徴量 [6]，SURF 特徴量 [7] ）と高レベルの特徴量（画像コンセプト・画像内容・画像構成）を採用する．なお，高レベルの特徴量のうち，画像コンセプトはデータセットに付随するラベルに従い，画像内容と画像構成は YOLO9000 [8] を用いて抽出する．このようにして得られる 6 種類の視覚特徴それぞれについて，画像

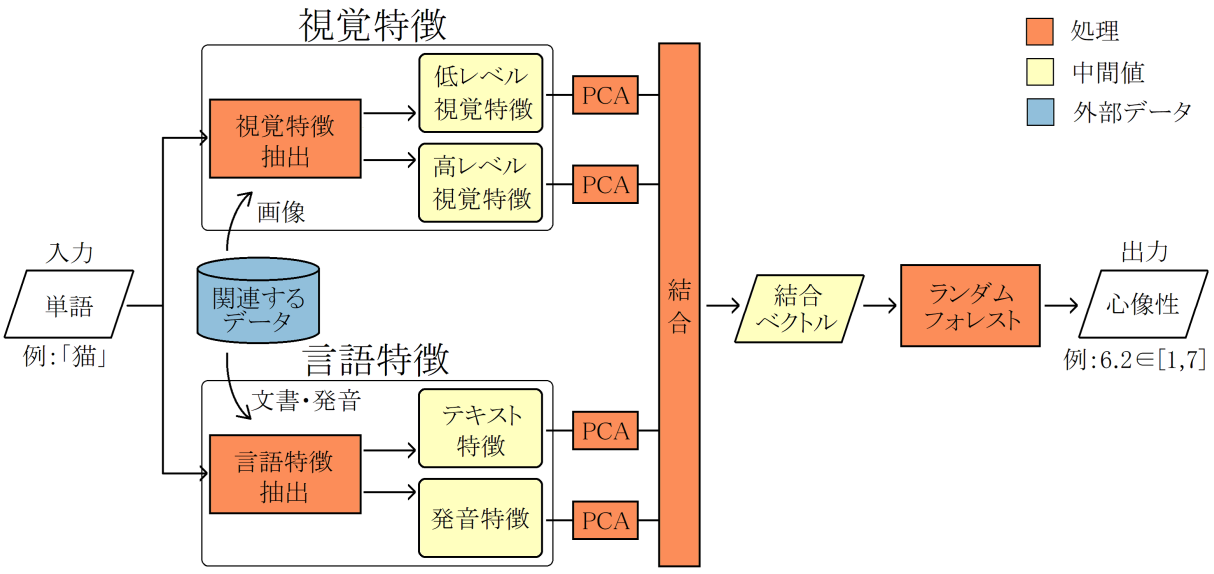

図 1: 提案手法の処理手順

間における類似度行列を求め，その固有値の上位 30 個を並べて特徴ベクトルを作成する．最後に，各レベルの特徴量ごとのこの特徴ベクトルを結合したものをそれぞれ低レベルの視覚特徴，高レベルの視覚特徴とする．

3.2 言語特徴の抽出

言語特徴として，テキスト特徴と発音特徴の 2 種類を考える．

テキスト特徴には，大規模なコーパスで事前学習を行った fastText [4] により得られる単語の分散表現を利用する．ここで，単語の分散表現は，その単語の文章中での位置取りや共起語などの情報を含んでいると考えられる．そこで本研究では，テキスト特徴として fastText により得られる分散表現をそのまま使用する．

次に，発音特徴の抽出について説明する．本研究では単語の発音を表現する手段として，国際音声学会が制定する国際音声記号（IPA）[1] を用いた単語の発音記号表記を使用する．

本研究では word2vec [9] と LSTM [10] を用いて発音特徴を抽出する．具体的には，word2vec 及び LSTM モデルにおいて，入力される文章と単語を，単語と音素に置き換えることでこれら（以降 pron2vec）を事前学習させる．ここで LSTM モデルの事前学習の真値には心像性を使用する．このようにして事前学習させた LSTM モデルの中間層の出力を発音特徴とする．発音特徴の抽出手順を図 2 に示す．なお，入力はデータセット中に現れる母音・子音・第 1 ストレス記号に限定し，他の記号は無視した．

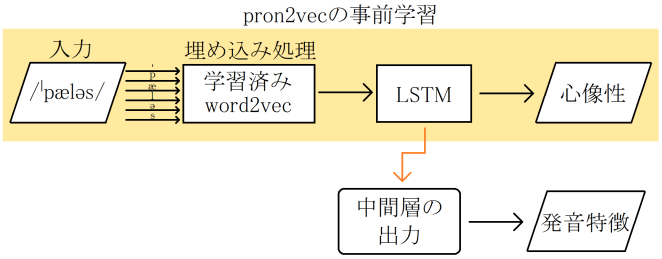

図 2: 発音特徴の抽出手順

3.3 特徴の統合と心像性の推定

上記手順により抽出した低レベルの視覚特徴，高レベルの視覚特徴，テキスト特徴，pron2vec による発音特徴の各々に対して，まず主成分分析（PCA）を適用し，同次元のベクトルに次元圧縮する．その後，それらのベクトルを結合し，ランダムフォレストにより回帰することで，心像性を推定する．

4 評価実験

提案手法により推定した英単語に対する心像性の評価実験を行なった．

[1] http://www.internationalphoneticassociation.org/content/ipa-chart/

4.1 データセット

本実験では回帰モデルの学習に使用する心像性データセット，画像データセット，発音データセット，発音特徴における pron2vec の事前学習に使用する心像性データセットの 4 種類のデータセットを使用した．

心像性のデータセットには Cortese らによるもの [11] と Reilly らによるもの [12] を併用した．これらは共に [1,7] の範囲の Likert 尺度で被験者による評定値を平均したものを単語の心像性としている．また，画像データセットには，画像共有サービス Flickr[2] に投稿された約 1 億枚の画像及び動画データにより構成される YFCC100M データセット [13] を使用した．

発音データセットの作成には，Macmillan Dictionary[3]を利用した．ここでは発音記法を統一するために，米国式の発音のみを採用し，第 1 ストレス記号が付与されていない 1 音節の単語に対しては，発音の先頭に第 1 ストレス記号を付与した．

pron2vec の事前学習には，回帰モデルの学習に使用したものと同一の心像性データセットを使用した．

4.2 実験概要

4.1 節のデータセットを用いて，587 語の英単語に対する心像性を推定した．視覚特徴としては低レベルの視覚特徴と高レベルの視覚特徴を併用した．テキスト特徴として使用する事前学習済みの分散表現として，Facebook 社から公開されている fastText[4] を使用した．

比較手法として，画像（視覚特徴）のみを用いた Kastner らの手法 [3] ，そして提案手法で用いる視覚特徴，言語特徴のうち一つの特徴のみを用いた手法を使用した．

4.3 実験結果

提案手法に対する評価実験の結果を表 1 に示す．視覚特徴・テキスト特徴・発音特徴を全て使用した提案手法において MAE が最小かつ相関係数が最大となり，精度が最も高い結果となった．これより，視覚特徴，テキスト特徴，発音特徴がそれぞれ異なる性質の情報を含んでいることが示唆された．

表 1: 提案手法による心像性推定結果．可読性を考慮し，心像性の範囲を [0,100] に変換して算出した．

使用した特徴量	MAE	相関係数
Kastner らの手法 [3]	10.14	0.63
視覚特徴	10.51	0.60
テキスト特徴	8.75	0.75
発音特徴	13.55	0.29
提案手法（全特徴）	**8.58**	**0.76**

さらに，Kastner ら [3] の手法と比較して，提案手法では MAE が約 17 %低下，相関係数が約 22 %向上することを確認した．一方で，提案手法において視覚特徴のみを使用した場合，Kastner らの手法と比較して推定精度が低下することも確認した．これは提案手法における特徴次元圧縮が原因であると考えられる．提案手法では低レベルの視覚特徴と高レベルの視覚特徴が同次元になるように特徴次元圧縮する．しかし，実際は低レベルの視覚特徴に対する最適な次元数と，高レベルの視覚特徴に対する最適な次元数が異なる可能性がある．そのため特徴次元圧縮手法の改良が必要であると言える．

最後に，いずれか 1 つの特徴を使用した比較手法及び全特徴を使用した提案手法について，評価用データの全単語に対する各手法の推定値，及び推定値に基づく最小 2 乗回帰の結果を図 3 に示す．全特徴を使用した提案手法による推定値が，テキスト特徴のみを用いた比較手法による推定値と類似する傾向が確認された．

5 おわりに

本発表では，視覚特徴と言語特徴を用いて単語の心像性を推定する手法を提案し，評価実験により推定した心像性の妥当性を確認した．

今後の課題として，特徴次元圧縮手法の改良，特徴量の追加，発音特徴における pron2vec の事前学習の改良などが挙げられる．

謝辞

本研究の一部は，科学研究費補助金及び国立情報学研究所との共同研究による．

[2] https://www.flickr.com/
[3] https://www.macmillandictionary.com/
[4] https://fasttext.cc/docs/en/pretrained-vectors.html

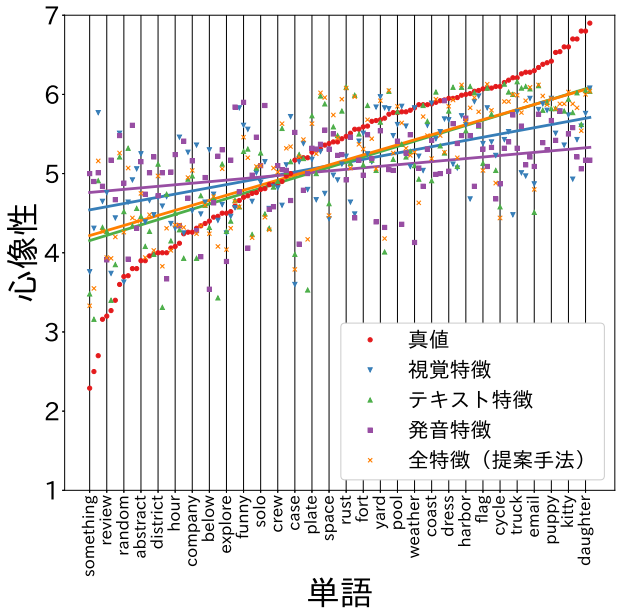

図 3: 使用特徴による推定値．[0,100] の範囲の心像性の推定値を [1,7] の Likert 尺度の範囲に変換した．

参考文献

[1] Allan Paivio, John C. Yuille, and Stephen A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.*, Vol. 76, No. 1, pp. 1–25, Jan. 1968.

[2] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proc. 2018 British Machine Vision Conf.*, No. 8, 14p., Sept. 2018.

[3] Marc A. Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimed. Tools Appl.*, 2020. Accepted for publication.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.*, Vol. 5, pp. 135–146, June 2017.

[5] Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proc. 56th Annual Meeting of the Assoc. for Comput. Linguist.*, pp. 217–222, July 2018.

[6] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of GIST descriptors for Web-scale image search. In *Proc. ACM Int. Conf. on Image and Video Retrieval 2009*, pp. 19:1–19:8, July 2009.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, Vol. 110, No. 3, pp. 346–359, June 2008.

[8] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *Proc. 2017 IEEE Conf. on Comput. Vision and Pattern Recogn.*, pp. 6517–6525, July 2017.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. 26th Int. Conf. on Neural Information Processing Systems*, Vol. 2, pp. 3111–3119, Dec. 2013.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, pp. 1735–1780, Dec. 1997.

[11] Michael J. Cortese and April Fugett. Imageability ratings for 3,000 monosyllabic words. *Beh. Res. Methods Instrum. Comput.*, Vol. 36, No. 3, pp. 384–387, Aug. 2004.

[12] Jamie Reilly and Jacob Kean. Formal distinctiveness of high-and low-imageability nouns: Analyses and theoretical implications. *Cogn. Sci.*, Vol. 31, No. 1, pp. 157–168, Feb. 2007.

[13] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Comm. ACM*, Vol. 59, No. 2, pp. 64–73, Feb. 2016.