# マルチモーダル情報を利用した単語ベクトル表現の生成

Learning word vectors using multimodal information

**内海　彰** (Utsumi AKIRA, Ph.D.)
**電気通信大学大学院 情報理工学研究科 教授**
(Professor, Department of Informatics, The University of Electro-Communications)
Cognitive Science Society　日本認知科学会　人工知能学会　言語処理学会　情報処理学会　日本心理学会
受賞：日本認知科学会大会発表賞 (1995)　人工知能学会研究分科会賞 (2006)　日本セキュリティ・マネジメント学会論文賞 (2011)
著書：人工知能と社会：2025 年の未来予想, オーム社 (2018)（共著）
メタファー研究 1, ひつじ書房 (2018)（共編著）
研究専門分野：認知科学　自然言語処理　人工知能

**あらまし**　単語の意味を数量化する手法として、意味空間モデルがある。意味空間モデルでは、大量の文章（コーパス）中の単語の共起情報を利用して、単語の意味を多次元ベクトルで表現する。しかし、単語の意味を言語情報だけから得るのは限界があるため、近年では、画像情報も利用したマルチモーダルな単語ベクトルの学習手法が研究されている。本研究では、人間が単語の意味を獲得する際のメカニズムを具現化するような新たなマルチモーダル単語ベクトル生成手法を提案する。提案モデルによる単語間類似度と人間の評定結果の相関を求めたところ、従来のマルチモーダル単語ベクトルに比べて、性能が向上することを確認した。さらに、コーパスなどの言語資源が整備されていない少資源言語に対して、画像を用いた意味ベクトルによる異言語間の対訳関係の推定を行い、その利用可能性を検証した。

## 1．はじめに

単語の意味を表現する手法として、意味空間モデル（もしくはベクトル空間モデル）が盛んに研究されている。意味空間モデルでは、コーパスと呼ばれる大量のテキストデータから、各単語の意味を多次元空間における特徴ベクトルとして表現し、単語間の意味的な類似性や関連性をそれらの単語ベクトルのコサインとして数量化する。自然言語処理の分野では、近年、深層学習やニューラルネットワークによる言語処理が主流になるにつれて、単語の意味表現としての単語ベクトルが事実上のデファクト・スタンダードになりつつある[1]。また、人間の認知メカニズムを解明しようとする認知科学の分野でも、人間の意味記憶の計算モデルとして広く使われている[2]。

しかし、我々人間がこの世に生まれてから単語の意味を学習していくときには、言語情報だけを用いているわけではない。特に、言語獲得がほとんど行われていない初期に学習する単語は、視覚などから得られる感覚情報を通じて学習される。よって、単語ベクトルの学習において、言語情報以外の非言語情報を考慮することは認知的に妥当であり、かつ、工学的にも興味深い方法である。このような背景から、図 1 に示すよ

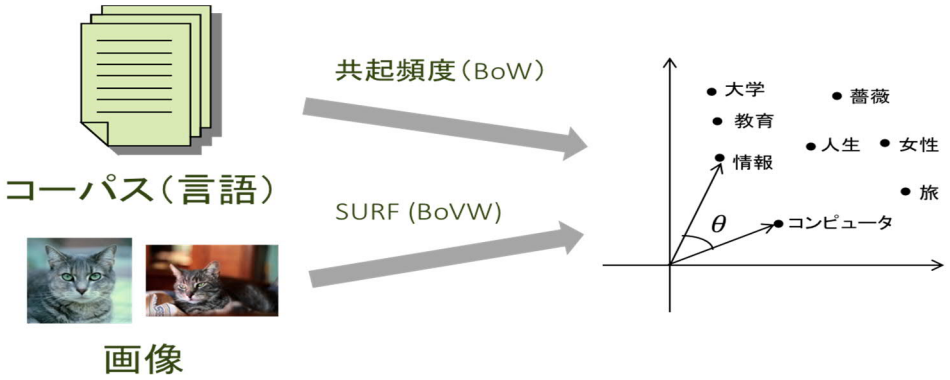


図 1　意味空間モデルによる単語の意味表現

うに、非言語情報、特に、画像から得られる視覚情報を加味したマルチモーダル単語ベクトルの学習手法が近年になって提案されている[3][4]。

これらの研究においては、言語情報のみから生成された単語ベクトルに比べて、マルチモーダル単語ベクトルの方が単語間の類似度判断などの性能が向上することが示されている。しかし、どのような単語についても有効なわけではなく、抽象語に対しては画像情報を考慮することによって、その表現性能が悪化することも示されている[5]。抽象語の意味を視覚などの感覚情報で表現するのが難しいことを考えると、人間の意味記憶のモデルとして単語ベクトルを考えたときに、この現象はごく自然である。そこで、本研究では、人間が抽象語の意味を学習するメカニズムをモデル化することによって、画像情報を考慮した単語ベクトルの性能向上を図ることを第一の目的とする。

一方で、画像情報は特定の言語に依存せずに言葉の意味を反映していることから、自然言語処理における画像情報の応用のひとつとして、多言語の単語を同一のベクトル空間で表現する多言語意味空間への適用が考えられる。言語コーパスだけから多言語意味空間を構築する研究は行われている[6][7]が、十分な規模の言語コーパスが入手できないような少資源言語に対しては、この手法を適用することができない。そこで、本研究では、画像情報を用いて少資源言語の単語ベクトルを生成する手法を提案し、その可能性を検討することを第二の目的とする。

## ２．間接的接地に基づくマルチモーダル単語ベクトルの生成

### ２．１ 背景

画像情報を用いた単語ベクトルで問題となるのが抽象語の扱いである。一般的に、抽象語に関しては典型的な画像イメージが存在しないため、抽象語に関する画像をそのまま用いて単語の画像ベクトルを計算するのは適切ではない。従来の手法[3] では、具象語と抽象語を区別せずに画像情報を用いているが、Kiela et al.[5] はこの問題を考慮し、抽象語に関しては単語の画像ベクトルを計算しないという手法を用いている。

しかし、記号接地の観点から見た場合、語彙の理解や獲得は本質的に知覚的な経験を通して得られるため、抽象語の意味表現に関しても知覚的な情報は考慮されるべきである。近年の記号接地に関する研究では、抽象概念の学習は具象概念を用いることで間接的に行われるという、間接的接地の概念が提案されている[8][9]。本研究では、この間接的接地の概念を取り入れた認知的に妥当な意味空間を構築し、記号接地モデルとしての妥当性を検証する。

### ２．２ 方法

本研究では、2.1 節で示した間接的接地の考え方を取り入れた意味空間を提案する。意味空間構築のアルゴリズムを以下に示す。

1. 文書コーパスを用いて意味空間$M_L$を生成する。
2. $M_L$に含まれる全単語$V$を具象語$V_C$と抽象語$V_A$に分類する。
3. 画像を用いた意味空間$M_V$を以下の手順で生成する。
   (a) 具象語 $V_C$ の各単語$w_i$に対しては、その画像情報を用いて画像特徴量ベクトル$\overrightarrow{v_i}$を生成する。
   (b) 抽象語$V_A$の各単語$w_i$に対しては、意味空間$M_L$によって計算された単語$w_i$の上位$N$件の近傍単語（意味的類似度が高い単語）から$V_C$に含まれる単語を$n$（$n < N$）個選択し、それらの画像特徴量ベクトルの重心を$w_i$の画像ベクトルとする。
4. 各単語に対して、意味空間$M_L$と$M_L$のベクトルを連結することによって、意味空間$M_G$を生成する。

上記の手順 2 の具象語と抽象語の分類には、Kiela et al.[5]が提案した単語の抽象度を画像から計算する手法を用いる。この手法は、抽象語は具象語に比べて画像の一貫性が低いという考えに基づいている。計算方法は、単語$w_i$の全画像ペアで画像特徴量ベクトルの距離を計算し、その距離の平均を単語$w_i$の抽象度とする。全単語を抽象度の降順によって並べ替え、抽象語の割合が$p_{abs}$となるように抽象度の高い単語を抽象語とする。

### ２．３ 評価実験

英語の単語に対して評価実験を行う。言語の意味空間$M_L$は、BNC（British National Corpus）を用いた共起頻度方式の特徴行列に対して非負相互情報量による重み付けを行い、特異値分解により 300 次元に圧縮して構築した。画像の意味空間$M_V$は次元数を 300 とし、SURF 特徴量を局所特徴量として用いて単語の画像ベクトルを計算し、非負相互情報量による重み付けを行った。意味空間に含まれる語彙数は、43,528 単語となった。

意味空間の評価には Wordsim-353 と SimLex-999 を用いた。これらは、人手で評定された英単語類似度評定データである。人手による類似度と意味空間から計算された類似度の相関係数によって、意味空間の性能を評価する。相関係数の値が高いほど、人間の意味記憶をより反映した意味空間であると言える。意味空間における単語どうしの類似度は、ベクトル間のコサイン類似度によって求める。比較対象の手法としては、テキストのみから生成された単語ベクトル、画像のみから生成された単語ベクトルとともに、従来のマルチモーダル単語ベクトルの 2 手法（単純に両方を用いる手法[3]と抽象語に対しては画像ベクトルを用いない手法[5]）を用いた。

### ２．４ 結果

評価結果を表 1 に示す。提案手法の精度（相関係数）が、その他の手法に比べて高くなっている。特に、従来のマルチモーダル単語ベクトル（テキスト＋画像の 2 手法）に比べて性能が向上している。これらの結果から、間接的接地に基づくマルチモーダル単語ベクトルの妥当性が示されたと言える。特に、提案手法の$p_{abs}$は 0.8 や 0.95 と非常に高い。このことから、画像から身体化が直接行われる具象語は少数であると言える。

しかし、その精度向上の程度は小さいため、間接的接地の概念を十分にモデル化しているとは言い難い。画像ベクトルを計算する元となる具象語の選択手法には、修正すべき点がいろいろと残されている。例えば、具象語をすべて対象とするのではなく、身体を通じた経験から直接形成される具象語を限定していく手法を考える必要がある。例えば、抽象語・抽象概念の形成には、感情が大きく関与することが実験的に明らかになっているので[10]、感情・情緒的に類似する具象語を選択することも、ひとつの方法として考えられるであろう。

## ３．画像ベクトルを用いた少資源言語の対訳語推定

### ３．１ 背景

画像は言語や文化に依存しない共通した意味を有する世界共通言語であると考えられる。例えば、「りんご」と「apple」は表層的には異なる単語であるが、両者は同じ意味を示していると言える。よって、画像のベクトル表現間の類似度を計算することで、異なる言語の単語間の意味的類似度の推定が可能である。本研究では、この画像の特徴を用いて、文書コーパスなどの言語資源が入手できないような言語、いわゆる少資源言語において、画像情報を用いた対訳語の推定を行う。具体的には、言語 A の単語集合と言語 B の単語集合が与えられたときに、対訳関係にある単語ペアを求める問題を画像ベクトルを用いて解くことを考える。

表１　間接的接地に基づく意味空間モデルによる単語間類似度データとの相関

| 意味空間モデル | WordSim-353 | SimLex-999 |
|---|---|---|
| テキストのみ $M_L$ | 0.525 | 0.248 |
| 画像のみ $M_V$ | 0.227 | 0.118 |
| テキスト＋画像（$M_L + M_V$）[3] | 0.476 | 0.235 |
| テキスト＋画像（抽象語は $M_L$ のみ）[5] | 0.532 | 0.245 |
| 提案手法 $M_G(P_{abs} = 0.95, n = 2, N = 30)$ | **0.551** | 0.260 |
| 提案手法 $M_G(P_{abs} = 0.80, n = 5, N = 30)$ | 0.535 | **0.278** |

### ３．２ 方法

言語 A の単語集合$S_A = \{a_i\}$、言語 B の単語集合$S_B = \{b_j\}$が与えられたときに、以下の方法で言語 A に対する言語 B の対訳語を推定する。

1. 単語集合$S_A$の画像ベクトル空間$V_A$と単語集合$S_B$の画像ベクトル空間$V_B$を、2.2 節で述べた方法で構築する。
2. 単語$a_i$の画像ベクトルを$v_i^A$、単語$b_j$の画像ベクトルを$v_j^B$とすると、次式を満たす対訳ペア$(a_i, b_k)$を求める。

$$b_k = argmax_j(\cos(v_i^A, v_j^B))$$

### ３．３ 評価実験

まず、Wikipedia の記事数などを参考に、表 2 に示す 13 言語を少資源言語として選定した。そして、使用頻度と具象度を考慮して英単語 100 語を選定し、各少資源言語に対して、これらの英単語の正解の対訳語を Google 翻訳により得た。さらに、英語と各少資源言語の対訳語をクエリとして、Bing 画像検索を行い、1 単語あたり最大で 30 枚の画像を取得した。

これらの実験データに対して、3.2 節で述べた方法で、各少資源言語に対する英語の対訳語を求めた。評価基準として、正しい対訳ペアが得られた割合を対訳率として求めた。加えて、3.2 節の方法は、類似度最大の単語のみを対訳語とする厳しい条件のため、類似度上位 10 件までを対訳語とした場合の対訳率も求めた。

### ３．４ 結果

対訳率の結果を表 3 に示す。全体的に、英語との対訳率はあまり高くない結果となった。このような結果が得られた理由の一つとして、少資源言語においては、単語の意味を適切に表す画像が取得できていないことが挙げられる。実際に、上記の問題が比較的生じないと思われる日本語で同様の実験を行った場合、類似度上位 10 件での対訳率が 0.55 となった。よって，画像検索の精度を上げることによって、本手法によって対訳語が推定できると言える。

また、本手法によって得られる対訳関係の性能が、言語間の比較言語学的な近さ（語族）と関係あるかどうかを検討するために、クラスタ分析を行った。2 言語間の距離を対訳関係にある単語の画像ベクトルどうしの距離（非類似度）の平均として、Ward 法による階層的クラスタリングを行った結果を図 2 に示す。

**表２　本研究で用いる少資源言語とそれらの言語族**

| 少資源言語 | 語 族 | 語 派 |
|---|---|---|
| ヨルバ語 | ニジェール・コンゴ語族 | |
| スンダ語 | オーストロネシア語族 | |
| アムハラ語 | アフロ・アジア語族 | |
| マケドニア語 | インド・ヨーロッパ語族 | スラブ語派 |
| イディッシュ語 | インド・ヨーロッパ語族 | ゲルマン語派 |
| ベンガル語 | インド・ヨーロッパ語族 | インド語派（東部語群） |
| グジャラト語 | インド・ヨーロッパ語族 | インド語派（中央語群） |
| パンジャブ語 | インド・ヨーロッパ語族 | インド語派（中央語群） |
| マラーティー語 | インド・ヨーロッパ語族 | インド語派（南部語群） |
| ネパール語 | インド・ヨーロッパ語族 | インド語派（パハール語群） |
| シンハラ語 | インド・ヨーロッパ語族 | インド語派（シンハラ・モルジブ語群） |
| スコットランド・ゲール語 | インド・ヨーロッパ語族 | ケルト語派 |
| アイルランド語 | インド・ヨーロッパ語族 | ケルト語派 |

表３　少資源言語と英語の対訳率

| 少資源言語 | 対訳率 | 上位 10 件までの対訳率 |
|---|---|---|
| アイルランド語 | 0.00 | 0.16 |
| アムハラ語 | 0.00 | 0.09 |
| イディッシュ語 | 0.00 | 0.13 |
| グジャラト語 | 0.01 | 0.19 |
| シンハラ語 | 0.03 | 0.19 |
| スコットランド・ゲール語 | 0.03 | 0.21 |
| スンダ語 | 0.02 | 0.19 |
| ネパール語 | 0.00 | 0.07 |
| パンシャブ語 | 0.04 | 0.18 |
| ベンガル語 | 0.02 | 0.20 |
| マケドニア語 | 0.06 | 0.26 |
| マラーティー語 | 0.00 | 0.19 |
| ヨルバ語 | 0.00 | 0.21 |
| 平均 | 0.02 | 0.17 |

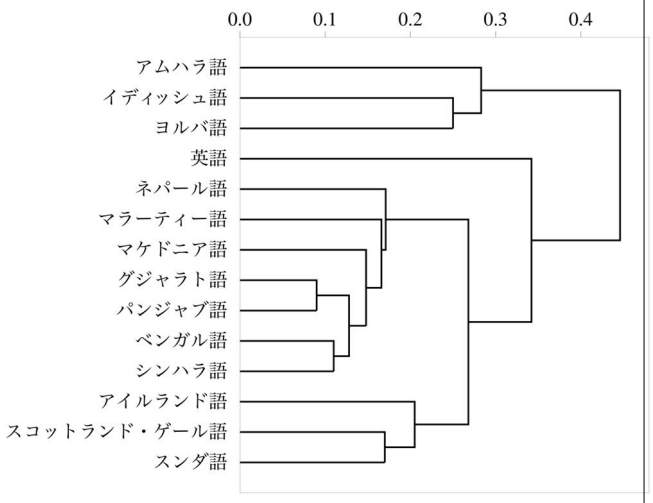


図２　画像ベクトルに基づく対訳関係から計算した類似度に基づく少資源言語のクラスタリング結果

その結果、インド・ヨーロッパ語族のうちのインド語派に属する 6 言語は、すべて同じクラスタに属しており、さらに、同じクラスタに分類される言語は言語族が近く、また、話される地域も近いという傾向が見られた。このことは、対訳語どうしの表す意味の内容や範囲が近いほど、画像ベクトルで計算する意味が近くなることを示しており、画像ベクトルによる単語の意味表現が有効であることの裏付けと考えられる。

## ４．おわりに

本研究では、単語の意味を表現する手法として意味空間モデルに注目し、画像情報を用いたマルチモーダル単語ベクトルの生成手法を新たに提案して、その有効性を確認するとともに、少資源言語の単語表現への応用可能性を示した。単語ベクトルに関する研究は、自然言語処理や認知科学の分野でも盛んに研究が行われており、新しい手法や知見が次々と得られている状況である。筆者の研究グループでは、本稿で紹介した研究以外にも、名詞句や文のベクトル表現や単語の語義ごとにベクトル表現を生成する手法の開発や、人間の意味記憶や意味・概念獲得の認知機構の解明など、単語ベクトルを対象とした様々な研究に取り組んでいる。また、単語ベクトル技術は、スポーツ記事などの専門的な文章の生成、小説を対象とした言語処理、会話エージェントなど、多岐にわたる言語処理への応用が可能である。今後も、引き続き単語ベクトルに注目して、これらの研究を推進していきたい。

## 参考文献

[1] Goldberg, Y., *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers (2017).

[2] Jones, M. N., Willits, J., & Dennis, S., Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford Handbook of Mathematical and Computational Psychology*, pp. 232-254, New York, NY: Oxford University Press (2015).

[3] Bruni, E., Tran, N. K., & Baroni, M., Multimodal distributional semantics. *Jounral of Artificial Intelligence Research*, Vol.49, pp.1-47 (2014).

[4] Silberer, C., Ferrari, V., & Lapata, M., Visually grounded meaning representations. *IEEE Transactions on Pattern Recognition and Machine Intelligence,* Vol.39, No.11, pp.2284–2297 (2017).

[5] Kiela, D., Hill, F., Korhonen, A., & Clark, S., Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 835-841 (2014).

[6] Utsumi, A., Multilingual distributional semantic models: Toward a computational model of the bilingual mental lexicon, In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science (EAP CogSci 2015)*, pp.270-275 (2015).

[7] 石渡 祥之佑, 鍛冶 伸裕, 吉永 直樹, 豊田 正史, 喜連川 優： 文脈語間の対訳関係を用いた単語の意味ベクトルの翻訳, 人工知能学会論文誌, Vol.31, No.6, AI30-A, pp.1-10 (2016).

[8] Louwerse, M. M., Symbol interdependency in symbolic and embodied cognition, *Topics in Cognitive Science*, Vol.3, pp. 273-302 (2011).

[9] Thill, S., Pado, S., & Ziemke, T., On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics, *Topics in Cognitive Science*, Vol.6, pp.545-558 (2014).

[10] Kousta, S.T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E., The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, Vol.140, pp.14–34 (2011).

この研究は、平成２６年度ＳＣＡＴ研究助成の対象として採用され、平成２７～２９年度に実施されたものです。