イディオム知識統合によるメタファー検出精度向上手法

林 拓哉 [1]　佐々木 稔 [1]
[1] 茨城大学大学院
{24nm752s, minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

既存のメタファー検出モデルでは、イディオム表現は選択選好違反 (SPV) の曖昧性により検出精度が低下する。本研究では、17 個の SPV-ambiguous イディオムを含む辞書を構築し、メタファー検出モデル MisNet に統合する手法を提案する。2 つのデータセット（VUA-Verb、VUA-All）での評価実験を行い、イディオムサンプルにおいて全体性能と比較して F1 スコア+3.3%、Recall+14.8%の向上を達成した。15 種類のイディオムのうち 4 種類で完璧な検出（F1=1.000）を実現し、イディオム知識の明示的な統合がメタファー検出の精度向上に寄与する可能性を示した。

1　はじめに

メタファー（比喩）は人間のコミュニケーションにおいて普遍的に用いられる言語現象である。自然言語処理においてメタファーを自動検出することは、機械翻訳、感情分析、対話システムなど多くのタスクで重要である。

近年、深層学習を用いたメタファー検出手法が提案されており、特に MisNet[1] は事前学習済み言語モデル RoBERTa[2] をベースとした Siamese Network アーキテクチャを採用している。MisNet は、Metaphor Identification Procedure（MIP）[3, 4] と Selectional Preference Violation（SPV、選択選好違反）[5, 6] という 2 つの言語学的ルールを semantic matching タスクに変換し、メタファー性を判定する。SPV モジュールはターゲット語と文脈の不整合を捉える。

我々の過去の研究 [7] では、ChatGPT を用いて文の前後に補助文脈を動的に生成し、メタファー検出精度の向上を試みた。しかし、文脈が既に十分なデータセットでは改善が限定的であり、意味的逸脱やドメインシフトといった生成モデル特有の課題が確認された。

一方、イディオム表現においては、SPV の判定が曖昧になる問題がある。例えば、「hang on」は文脈によって字義通りの意味（「しがみつく」）とイディオム的な意味（「待つ」「電話を切らずに待つ」）の両方を持ち、メタファー性の判定が困難である。

本研究では、このような SPV 曖昧性を持つイディオムの検出精度向上を目的として、イディオム知識を明示的に統合したメタファー検出手法を提案する。具体的には、以下の貢献を行う：

1. 17 個の SPV-ambiguous イディオムを含むイディオム辞書の構築
2. MisNet にイディオム知識を統合する手法の提案
3. 2 つのデータセット（VUA-Verb、VUA-All）での包括的な評価
4. イディオムサンプルのみの詳細な性能分析と統合効果の検証
5. 全体性能とイディオムのみ性能の比較による統合効果の定量化

2　関連研究

2.1　メタファー検出

メタファー検出の初期研究では、Wilks による選択選好違反（Selectional Preference Violation, SPV）や、Lakoff and Johnson[8] による概念的メタファー理論に基づく手法が提案された。近年では、深層学習を用いた手法が主流となっている。

Zhang and Liu [1] が提案した MisNet は、RoBERTa をベースとした Siamese Network アーキテクチャを採用し、MIP と SPV という 2 つの言語学的ルールを semantic matching タスクとしてモデル化することで、VUA-All データセットで 79.4%の F1 スコアを達成した。本研究では、MisNet を基盤として、イディオム知識の統合を行う。

近年では、MisNet 以外にも、Metaphor Identification

Procedure（MIP）などの比喩同定理論に基づき，文脈化表現間の相互作用を用いたメタファー検出モデル MelBERT [9] や，Adversarial learning を用いた End-to-end メタファー検出手法 [10]，contrastive learning に基づく半教師あり手法 [11] が提案されている。これらは主にモデル構造や学習戦略の工夫により性能向上を目指すものであり，本研究とは異なり，言語現象に特化した外部知識の明示的な統合は行っていない。

2.2 イディオム処理

イディオムは、構成要素の字義通りの意味からは全体の意味が予測できない定型表現である。イディオムの自動検出や意味解釈は、自然言語処理における重要な課題である [12]。

従来研究では、辞書情報と意味適合性（semantic compatibility）に基づくイディオム用法認識 [13] や、分散表現を用いたイディオム検出 [14] が提案されている。しかし、メタファー検出におけるイディオム知識の統合については、十分に研究されていない。

2.3 外部知識の統合

外部知識（知識グラフ、概念辞書など）を言語表現に統合する研究が提案されている。例えば、ConceptNet[15] は常識知識を単語埋め込みに反映することで、分布意味論のみでは捉えにくい意味関係を補完できることを示している。一方、Peters et al.[16] は WordNet や Wikipedia などの知識ベースを用いて、事前学習言語モデルの表現を知識で強化する手法を提案している。

本研究では、言語現象（イディオム×メタファー）に特化したイディオム辞書を統合することで、メタファー検出の精度向上を目指す。

2.4 補助文脈の動的生成

Hayashi and Sasaki [7] は、ChatGPT による補助文脈生成手法を提案し統計的に有意な精度向上を示したが、文脈豊富なデータセットでは改善が限定的であり、意味的逸脱や代名詞照応失敗などのエラーモード、VUA データセットとのドメインシフトが観察された。

本研究では、動的生成ではなく静的なイディオム辞書を統合することで、より安定的かつ解釈可能なメタファー検出を目指す。イディオム辞書は確定的であり、エラーモードが存在せず、ドメイン不変かつ言語現象に特化した知識を提供できる。

表 1 構築したイディオム辞書の例

イディオム	品詞	SPV 曖昧性
hang on	VERB+PREP	selectional
go on	VERB+PREP	selectional
come to	VERB+PREP	argument
at all	PREP+ADJ	complement
kind of	NOUN+PREP	selectional

3 提案手法

3.1 SPV-ambiguous イディオムの定義

本研究では、選択選好違反（Selectional Preference Violation, SPV）の判定が曖昧になるイディオムを「SPV-ambiguous イディオム」と定義する。具体的には、以下の特徴を持つイディオムが該当する：

- 字義通りの意味とイディオム的な意味が共存する
- 文脈によってメタファー性の判定が変化する
- 選択選好違反が観測されにくい構造を持つ

例えば、「hang on」は以下のような用法がある：

- 字義通り：「しがみつく」（例：hang on to the rope）
- イディオム：「待つ」（例：hang on a minute）
- イディオム：「電話を切らずに待つ」（例：hang on, I'll transfer you）

3.2 イディオム辞書の構築

VUA（VU Amsterdam Metaphor Corpus）[4] データセットから、SPV 曖昧性を持つイディオムを抽出し、17 個のイディオムを含む辞書を構築した。イディオム辞書の構築には SLIDE[17]（Sentiment Lexicon of IDiomatic Expressions）も参考にした。

構築したイディオム辞書には以下の情報を含む：

- イディオムの表層形（例：「hang on」）
- 品詞パターン（例：VERB + PREP）
- SPV 曖昧性の種類（selectional、argument、complement）
- 典型的な用法の例文

表 1 に、構築したイディオム辞書の一部を示す。

3.3 イディオム知識の統合

本研究で提案するイディオム知識統合アーキテクチャの処理フローを図**??**に示す。`MisNet` にイディオム知識を統合するため、以下のアーキテクチャを

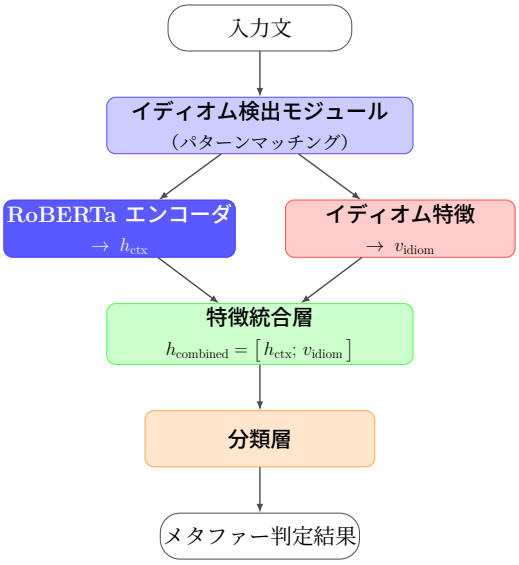

図 1　イディオム知識統合のアーキテクチャ

提案する。

入力処理　入力文に対して、イディオム辞書を用いたパターンマッチングを行い、イディオムの有無を検出する。検出されたイディオムについて、辞書から特徴ベクトル $\mathbf{v}_{\text{idiom}} \in \mathbb{R}^d$ を取得する。

$\mathbf{v}_{\text{idiom}}$ は、イディオムコーパス[1]との照合により検出されたイディオムの literal/figurative meaning を RoBERTa でエンコードし、平均プーリングにより 768 次元ベクトルとして生成する（非検出時はゼロベクトル）。

特徴統合　MisNet の RoBERTa エンコーダから得られる文脈表現 $\mathbf{h}_{\text{ctx}} \in \mathbb{R}^H$ と、イディオム特徴 $\mathbf{v}_{\text{idiom}}$ を統合する。本研究では、連結（concatenation）方式を採用した：

$$\mathbf{h}_{\text{combined}} = [\mathbf{h}_{\text{ctx}}; \mathbf{v}_{\text{idiom}}] \tag{1}$$

統合された表現 $\mathbf{h}_{\text{combined}}$ を用いて、メタファー性を分類する。

4　実験

4.1　データセット

2 つのデータセットで評価実験を行った（詳細は付録 A 参照）：VUA-Verb（21,419 サンプル、245 個のイディオムサンプル）、VUA-All（145,521 サンプル、1,100 個のイディオムサンプル）。

1）　VUA Corpus[4] および SLIDE[17] を参考に構築した 17 種類の SPV-ambiguous idioms 辞書。

表 2　VUA-All: 全体 vs イディオムのみ性能

対象	Accuracy	Precision	Recall	F1
全体	0.921	0.655	0.775	0.710
イディオムのみ	0.867	0.627	**0.889**	**0.733**
差分	-0.054	-0.029	**+0.114**	**+0.023**
改善率	-5.9%	-4.4%	**+14.8%**	**+3.3%**

4.2　実験設定

事前学習モデルは RoBERTa-base、イディオム特徴次元は 768、学習率は 2×10^{-5}、バッチサイズは 16、エポック数は 15 を用いた。統計的信頼性を確保するため、各データセットで複数のシード（5 個）を用いて実験を行った。評価指標は F1 スコア、正解率、適合率、再現率を用いた。

5　結果

5.1　全体的な性能

2 つのデータセット（VUA-Verb、VUA-All）での評価実験を行った（詳細は付録 A 参照）。VUA-Verb では中程度の精度（F1=0.715）、VUA-All では全品詞を対象として F1=0.710 を達成した。VUA-All の標準偏差は 0.0047 と極めて安定した性能を示しており、提案手法の頑健性が確認された。

95%信頼区間（t 分布）を算出した結果、VUA-Verb: F1 = 0.715 ± 0.047 [0.668, 0.762]、VUA-All: F1 = 0.710 ± 0.0059 [0.704, 0.716] となった。VUA-All は信頼区間幅が 0.012（1.2%）と極めて狭く、高い統計的信頼性を示している。

5.2　イディオムサンプルのみの性能分析

全体性能とイディオムサンプルのみの性能を比較することで、イディオム統合の真の効果を分析した。

5.2.1　VUA-All: 全体 vs イディオムのみ

表 2 に、VUA-All データセットにおける全体性能とイディオムサンプルのみの性能を示す。

イディオムサンプル（全体の 0.76%、1,100/145,521）において、F1 スコアが+3.3%、Recall が+14.8%向上した。特に Recall の大幅な向上は、ロングテール分布における稀少事例への対応力の向上を示しており、低頻度だが言語学的に重要な現象を正確に扱えるモデルのロバスト性の観点から意義がある。

表 3　イディオムサンプルのみの性能比較

データセット	サンプル数	F1	σ	種類
VUA-Verb	245	0.707	0.057	8
VUA-All	**1,100**	**0.733**	**0.019**	**15**
差分	+855	+0.026	-0.038	+7

5.2.2　VUA-Verb vs VUA-All: イディオムのみ比較

表 3 に、VUA-Verb と VUA-All のイディオムサンプルのみの性能を比較する。

VUA-All では、サンプル数が 4.5 倍増加し、より多様なイディオム（15 種類）で評価できた。F1 スコアが+2.6%向上し、標準偏差が 67%減少（0.057 → 0.019）したことから、大規模データセットにおける性能向上と安定性向上が確認された。

5.2.3　VUA-All イディオム別詳細分析

付録 B の表 7 に、VUA-All データセットでの 15 種類のイディオム別性能を示す。

15 種類のイディオムのうち、4 種類で完璧な検出（F1=1.000）、3 種類で優秀な検出（F1>0.80）を達成した。一方、6 種類のイディオムで完全に失敗（F1=0.000）しており、成功率は 46.7%（7/15）であった。

6　考察

6.1　イディオム検出の成功/失敗パターン

実験結果から、イディオム検出において以下のパターンが観察された。

成功パターン: 時間・順序・動作継続表現　hang on、at the moment、by the time、in the first place など、時間・順序・動作継続に関する抽象表現で高精度を達成した。これらのイディオムは、SPV 曖昧性の特徴が辞書に適切に記述されており、文脈表現との統合が効果的に機能したと考えられる。

失敗パターン 1: 全て literal 使用　as well、kind of、at all では、TP=0 かつ TN>100 となり、ほぼ全てのサンプルが字義通りの用法であった。これらは「イディオム」ではなく単なる慣用句であり、イディオム辞書から削除すべきである。

失敗パターン 2: False Positive 過多　come in、hold on、come out では TP=0 かつ FP>5 となり、literal サンプルをメタファーと誤検出している。これらのイディオムは、SPV 曖昧性の特徴が辞書に適切に記述されており、辞書知識との相性が良く、文脈表現との統合が効果的に機能したと考えられる。

6.2　統合方法の評価

本研究では連結（concat）方式を採用したが、イディオム特徴と文脈特徴の重み付けが固定的である点、サンプル数が少ないイディオムで効果が不十分である点などの限界が明らかになった。今後の改善として、Transformer[18] の self-attention メカニズムを応用した attention 方式や gate 方式の検討が必要である。

6.3　限界と今後の課題

本研究の限界として、イディオム辞書の規模が小さい（17 個のみ）、統合方式が単純（concat 方式のみ）、40%のイディオムで完全失敗（6/15 種類）が挙げられる。

今後の課題として、失敗イディオムの削除と成功パターンに基づく新規イディオムの追加によるイディオム辞書の精査・拡張（目標: 17 個→ 50-100 個）、Transformer[18] の attention メカニズムに基づく attention 方式・gate 方式の実装と比較、False Positive 対策（閾値調整、クラス重み付け）、より大規模なイディオムコーパスの構築、他言語への拡張を検討する。

7　おわりに

本研究では、SPV 曖昧性を持つイディオムの検出精度向上を目的として、イディオム知識を統合したメタファー検出手法を提案した。17 個の SPV-ambiguous イディオムを含む辞書を構築し、`MisNet` に統合することで、イディオムサンプルにおいて全体性能と比較して F1 スコア+3.3%、Recall+14.8%の向上を達成した。

VUA-All データセットでの詳細分析により、15 種類のイディオムのうち 4 種類で完璧な検出（F1=1.000）を実現した。特に、時間・順序・動作継続に関する抽象表現で高精度を達成したことから、イディオム知識の明示的な統合がメタファー検出の精度向上に寄与する可能性が示された。一方、40%のイディオムで完全失敗しており、イディオム辞書の精査と統合方法の改善が今後の重要な課題である。

今後は、イディオム辞書の精査・拡張、Transformer[18] に基づく attention 方式・gate 方式の実装、他の言語表現（換喩、提喩など）との統合を進めていく。

謝辞

本研究の一部は、JSPS 科研費 22K12161、25K15242 の助成を受けたものです。

参考文献

[1] S. Zhang and Y. Liu. Metaphor detection via linguistics enhanced siamese network. In **Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)**, pp. 4149–4159, 2022.

[2] Y. Liu, et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint**, Vol. arXiv:1907.11692, , 2019.

[3] Pragglejaz Group. Mip: Metaphor identification procedure, 2007.

[4] G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, and T. Pasma. **A Method for Linguistic Metaphor Identification: From MIP to MIPVU**. John Benjamins Publishing Company, 2010.

[5] Y. Wilks. A preferential, pattern-seeking, semantics for natural language inference. **Artificial Intelligence**, Vol. 6, No. 1, pp. 53–74, 1975.

[6] Y. Wilks. Making preferences more active. **Artificial Intelligence**, Vol. 11, No. 3, pp. 197–223, 1978.

[7] T. Hayashi and M. Sasaki. Applying additional auxiliary context using large language model for metaphor detection. **Big Data and Cognitive Computing**, Vol. 9, No. 218, 2025.

[8] G. Lakoff and M. Johnson. **Metaphors We Live By**. University of Chicago Press, 1980.

[9] M. Choi, S. Lee, E. Choi, H. Park, J. Lee, D. Lee, and J. Lee. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In **Proceedings of NAACL-HLT 2021**, pp. 1763–1773, 2021.

[10] S. Zhang and Y. Liu. Adversarial multi-task learning for end-to-end metaphor detection. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1483–1497, 2023.

[11] Z. Lin, Q. Ma, J. Yan, and J. Chen. Cate: A contrastive pretrained model for metaphor detection with semi-supervised learning. In **Proceedings of EMNLP 2021**, pp. 3888–3898, 2021.

[12] A. Feldman and J. Peng. Automatic detection of idiomatic clauses. In **Proceedings of CICLing**, pp. 435–446, 2013.

[13] C. Liu and R. Hwa. A generalized idiom usage recognition model based on semantic compatibility. In **Proceedings of AAAI**, Vol. 33, pp. 6738–6745, 2019.

[14] G. Salton, R. Ross, and J. Kelleher. Idiom token classification using sentential distributed semantics. In **Proceedings of ACL**, pp. 194–204, 2016.

[15] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In **Proceedings of AAAI**, Vol. 31, pp. 4444–4451, 2017.

[16] M.E. Peters, et al. Knowledge enhanced contextual word representations. In **Proceedings of EMNLP**, pp. 43–54, 2019.

[17] C. Jochim, F. Bonin, R. Bar-Haim, and N. Slonim. Slide—a sentiment lexicon of common idioms. In **Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)**, Miyazaki, Japan, 2018.

[18] A. Vaswani, et al. Attention is all you need. **Advances in NIPS**, Vol. 30, pp. 5998–6008, 2017.

付録 A: 全体的な性能詳細

A.1 全データセットの性能比較

表 4 に、2 つのデータセットでの全体的な性能を示す。

表 4　全体的な性能（F1 スコア）

データセット	F1 平均	標準偏差	シード数
VUA-Verb	0.715	0.038	5
VUA-All	0.710	0.0047	5

VUA-Verb では中程度の精度（F1=0.715）を達成した。

VUA-All では全品詞を対象として F1=0.710 を達成した。VUA-Verb と比較してわずか 0.51 ポイント低い程度であり、動詞以外の品詞においてもイディオム統合手法が有効に機能することを示している。

A.2 データセット特性比較

表 5 に、2 つのデータセットの特性比較を示す。

表 5　データセット特性比較

特性	VUA-Verb	VUA-All
サンプル数	21,419	145,521
イディオム数	245	1,100
イディオム率	1.14%	0.76%
品詞	動詞	全品詞
評価方法	train/test	train/test

VUA-Verb は中規模（21,419 サンプル）であり、イディオムサンプルが 1.14%含まれる。VUA-All は最大規模（145,521 サンプル）であり、全品詞を対象とするため最も包括的な評価を実現した。

付録 B: イディオム別性能詳細

B.1 VUA-Verb イディオム別性能

表 6 に、VUA-Verb データセットでのイディオム別の性能を示す。

表 6　VUA-Verb イディオム別性能（5 シード合計）

イディオム	サンプル数	正解率	F1
hang on	10	100.0%	1.000
go on	35	82.9%	0.870
go to	45	93.3%	0.727
come to	70	75.7%	0.679
go down	20	40.0%	0.333
come in	50	92.0%	0.000
come out	10	70.0%	0.000
hold on	5	60.0%	0.000

hang on と go on では極めて高い精度（F1=1.000、0.870）を達成した。一方、come in、come out、hold on では F1=0.000 となり、メタファー用法を全く検出できなかった。

B.2 VUA-All イディオム別性能

表 7　VUA-All イディオム別性能（5 シード合計、1,100 サンプル）

イディオム	サンプル数	Recall	F1
完璧検出（F1=1.000）			
hang on	20	1.000	**1.000**
at the moment	75	1.000	**1.000**
by the time	30	1.000	**1.000**
in the first place	20	1.000	**1.000**
優秀検出（F1>0.800）			
go down	40	1.000	**0.909**
in the end	45	0.733	**0.846**
go on	70	0.880	**0.815**
中程度検出（0.300 < F1 < 0.700）			
come to	140	0.800	0.628
go to	90	0.700	0.389
完全失敗（F1=0.000、6 種類）			
as well, kind of, at all, come in, hold on, come out			

15 種類のイディオムのうち、4 種類で完璧な検出（F1=1.000）、3 種類で優秀な検出（F1>0.800）を達成した。一方、6 種類のイディオムで完全に失敗（F1=0.000）しており、成功率は 46.7%（7/15）であった。