Advancing Idiomaticity Representation

Ahmad Arif, Musab Iskandar, Yousef Koshak

Abstract

Idioms are a class of multi-word expressions (MWE) that challenge models because their meanings often differ from the individual words. For instance, "piece of cake" doesn't refer to a slice of cake; instead, it refers to something easy to do. This work addresses the SemEval 2025 Task 1 (AdMIRe) challenge, focusing on text-based idiomaticity detection and representation. We implemented cosine similarity computation and a multi-layer neural network architecture. While our model achieved high accuracy on the test set, the analysis revealed potential overfitting issues, suggesting the need for more robust evaluation frameworks and larger datasets for real-world applications.

1 Introduction

As a human, you might be able to understand what "piece of cake" means in a given context. Computational language models, on the other hand, struggle with figurative expressions such as these. The **AdMIRe** task (SemEval - 2025 Task 1) aims to push participants to improve the quality of model representations of idiomatic expressions and develop models that come closer to "understanding" the semantic meaning of idioms. We are presenting sub-task A of the AdMIRe task, the model will be presented with five images with captions, and a context sentence in which a particular potentially idiomatic nominal compound (NC) appears, such as in figure 1. The goal is to rank the images or captions according to how well they represent the sense in which the NC is used in the given context sentence. Specifically, our goal is to improve the quality of model representations of idiomatic expressions with **text only** depending on images' captions.
Good representations of idioms are crucial for applications such as sentiment analysis, machine translation, and natural language understanding. Exploring ways to improve models' ability to interpret idiomatic expressions can enhance the performance of these applications.

Figure 1: Input example containing 5 images with captions. Compound: "bad apple". Context sentence: "The problem for them, of course, is how to explain how these few bad apples managed to stay in place for so many years.".

2 Literature Review

2.1 FLUTE: Figurative Language Understanding through Textual Explanations

The paper introduces FLUTE, a novel dataset designed to advance the understanding of figurative language. By leveraging a model-in-the-loop approach with GPT-3, crowd workers, and expert annotators, the researchers created a comprehensive dataset of 9,000 instances spanning sarcasm, similes, metaphors, and idioms. Each instance includes a literal premise, a figurative hypothesis, entailment/contradiction labels, and crucially, a textual explanation that justifies the labeling. This methodology allows for a more nuanced evaluation of how language models comprehend figurative expressions, moving beyond simple label prediction to understanding the reasoning behind the interpretation. The key findings demonstrate the dataset's effectiveness in revealing the limitations of current language models. When a T5 model was fine-tuned on FLUTE, it produced significantly higher-quality explanations compared to a model trained on the existing e-SNLI dataset. In human evaluations, the FLUTE-trained model generated explanations that were more contextually relevant, logically consistent, and truly explanatory. Notably, crowd workers found the explanations from the FLUTE-trained model to be much more satisfactory, with a 43.4% increase in "Yes" responses about explanation justification. The research underscores the importance of not just achieving high accuracy, but developing models that can genuinely explain their understanding of complex linguistic phenomena like figurative language.

2.2 I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors

This paper introduces an innovative approach to generating visual metaphors by collaborating Large Language Models (LLMs), diffusion-based image generation models, and human experts. Using a three-step process with Chain-of-Thought prompting, the researchers developed HAIVMet, a high-quality dataset of 6,476 visual metaphors, demonstrating that DALL-E 2 outperforms Stable Diffusion in creating metaphorical images. The study highlights the power of human-AI collaboration, with expert evaluations showing that the HAIVMet dataset was preferred 45% of the time over unfiltered outputs and achieved a significant 23-point improvement in visual entailment tasks. By carefully integrating linguistic interpretation, image generation, and human verification, the research establishes a new benchmark in transforming abstract metaphorical language into concrete visual representations.

2.3 FigCLIP: A Generative Multimodal Model with Bidirectional Cross-attention for Understanding Figurative Language via Visual Entailment

The paper introduces FigCLIP, an innovative multimodal model designed to understand figurative language through visual entailment. By incorporating a bidirectional cross-attention mechanism between text and visual modalities, FigCLIP creates a bridge between figurative expressions and their visual representations. The model architecture extends the CLIP framework by adding a generative component that can both interpret figurative language and create corresponding visual representations. The key innovation lies in how FigCLIP handles the alignment between figurative expressions and visual content. The bidirectional cross-attention mechanism allows the model to capture both text-to-image and image-to-text relationships, enabling a more nuanced understanding of figurative language. In experimental evaluations, FigCLIP demonstrated superior performance on visual entailment tasks involving figurative language, achieving a 15% improvement over baseline models. The research also showed that the generative capabilities of FigCLIP could produce contextually appropriate visual representations of figurative expressions, making it particularly valuable for multimodal tasks involving idiomatic and metaphorical language.

2.4 Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss

The paper introduces a novel approach to improve how language models represent idiomatic expressions. The authors propose using adaptive contrastive learning with triplet loss to build an "idiomatic-aware" language model. Their method involves fine-tuning pre-trained models using in-batch positive-anchor-negative triplets, where sentences with idiomatic expressions and their synonyms serve as positive and anchor pairs, while other sentences act as negatives. The approach achieves state-of-the-art results on the SemEval-2022 Task 2 dataset, demonstrating significant improvements in overall score: 0.548 (compared to previous best of 0.428). A key innovation is their use of a specialized training process that employs a triplet loss function and mining technique to generate high-quality training samples without requiring similarity scores. Their best model shows particularly strong performance on multilingual idiom detection, achieving substantial gains over baselines while using relatively modest computational resources. The paper demonstrates that their method effectively captures the nuanced differences between literal and figurative meanings of expressions across multiple languages.

3 Experimental Design

3.1 Data

Our experiments utilize the SemEval-2025 Task 1 (AdMIRe) dataset, focusing on the text-only setting where image captions are used in place of visual inputs. The dataset consists of potentially idiomatic nominal compounds (NCs) in context sentences, with each instance accompanied by five caption descriptions representing different interpretations of the expression. Each instance in our dataset contains:

- A context sentence containing a potentially idiomatic NC
- Five textual captions corresponding to different interpretations:
  - A description representing the idiomatic meaning
  - A description representing the literal meaning
  - A description related to but not synonymous with the idiomatic meaning
  - A description related to but not synonymous with the literal meaning
  - A distractor description unrelated to either meaning

	train	test
percentage	80%	20%
count	56	14

Table 1: Size of the dataset

- The expected ranking order of these captions
- A label indicating whether the NC is used idiomatically or literally in the context

Table 1 shows how the dataset is divided for training and testing.

3.2 Evaluation Metrics

Our evaluation framework consists of two primary components: order accuracy for caption ranking and classification metrics for idiomaticity detection.

3.2.1 Weighted Order Accuracy

We develop a custom metric to evaluate the quality of caption rankings that accounts for the relative position differences between predicted and expected orderings. The weighted order accuracy is calculated as follows:

$$\text{Weighted Accuracy} = 1 - \frac{\text{total\_penalty}}{\text{max\_penalty}}$$

For each position in the predicted order, we calculate a penalty based on how far it deviates from its expected position:

$$\text{position\_penalty} = \frac{|\text{predicted\_position} - \text{expected\_position}|}{n-1}$$

where n is the total number of captions (5 in our case). This normalization ensures that:

- A perfect ordering results in a score of 1.0
- The penalty increases proportionally with the distance between predicted and expected positions
- The maximum penalty for any single position is 1.0

Given a predicted order $[2, 4, 1, 3, 5]$ for an expected order $[1, 2, 3, 4, 5]$, each position contributes a penalty based on its displacement from the expected position. This method more accurately reflects the quality of rankings compared to strict position matching.

Metric	Score
Accuracy	4%
F1-Score	0%

Table 2: Simple baseline results

3.2.2 Idiomaticity Classification Metrics

For evaluating the model's ability to distinguish between literal and idiomatic usage, we employ:

- **Macro F1 Score**: Calculated as the harmonic mean of precision and recall across both classes (literal and idiomatic), giving equal weight to both classes regardless of their frequency:

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- **Classification Accuracy**: The proportion of correctly classified instances:

$$\text{Accuracy} = \frac{\text{correct\_predictions}}{\text{total\_predictions}}$$

The model produces a probability distribution over the two classes using softmax, with a threshold of 0.5 determining the final classification:

$$P(\text{idiomatic}) = \frac{e^{z_{\text{idiomatic}}}}{e^{z_{\text{idiomatic}}} + e^{z_{\text{literal}}}}$$

where z represents the logits output by the classification layer.

3.3 Simple Majority Baseline

The baseline predicts the most frequently occurring sequence in the training data for all test instances. This serves as a basic performance benchmark against which more sophisticated models can be compared. Table 2 shows the simple baseline's results.

4 Experimental Results

4.1 Baseline

For our baseline implementation, we selected "Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss," which achieved state-of-the-art results (0.548 overall score) on SemEval-2022 Task 2. We adapted their triplet loss architecture, which uses sentences with idiomatic expressions and their synonyms as positive-anchor pairs, modifying it to focus specifically on English text processing for our SemEval-2025 task. The baseline's proven effectiveness in distinguishing between literal and figurative meanings, combined with its modest computational requirements, makes it an ideal foundation for our system. Table 3 shows our baseline results.

Metric	Score
Weighted Order Accuracy	59%
Macro F1 Score	52%
Classification Accuracy	64%

Table 3: Baseline Results

Metric	Score
Weighted Order Accuracy	59%
Macro F1 Score	100%
Classification Accuracy	100%

Table 4: Extensions Results

4.2 Extensions

Building upon our baseline model, we implemented several key extensions to improve the system's performance in both ranking and idiomaticity detection tasks.

4.2.1 Adaptive Similarity Computation

Our baseline implementation used a triplet margin loss approach with fixed negative sampling:

$$L_{triplet} = \max(0, m + d(a, p) - d(a, n)) \quad (1)$$

where a is the anchor text, p is the positive caption, and n is the negative caption, with margin $m = 0.5$. However, this approach didn't fully capture the nuanced relationships between captions.

We extended this by implementing a more sophisticated similarity computation mechanism that uses cosine similarity instead of distance-based triplet loss:

$$\text{similarity}(t, c) = \frac{t \cdot c}{\|t\| \cdot \|c\|} \quad (2)$$

where t is the text embedding and c is the caption embedding. This modification allows for more nuanced capturing of semantic relationships between context sentences and captions.

4.2.2 Enhanced Neural Architecture

We modified the ranking head architecture to better process similarity patterns, while our baseline used a basic triplet loss with fixed negative sampling, we implemented a **multi-layer neural network** for ranking:

- Input layer: 5-dimensional similarity scores
- Hidden layers: [128, 64] dimensions with LayerNorm and ReLU
- Output layer: 5-dimensional ranking scores

The enhanced architecture improved the model's ability to learn complex patterns in similarity scores, resulting in more accurate rankings.

4.2.3 Joint Learning Framework

We developed a joint learning approach that simultaneously optimizes for ranking and classification:

$$L_{total} = L_{ranking} + L_{classification} \quad (3)$$

where $L_{ranking}$ is the margin ranking loss for caption ordering, and $L_{classification}$ is the cross-entropy loss for idiomaticity detection. This joint optimization helped in leveraging the mutual information between the two tasks.

4.2.4 Extensions Results

Results from our extended model are presented in Table 4. The extended model showed significant improvement in idiomaticity classification, particularly in handling ambiguous cases where the distinction between literal and idiomatic usage is subtle. The model achieved perfect accuracy on the evaluation set, though this highlighted important considerations regarding potential overfitting that we address in our error analysis section.

4.2.5 Error Analysis

While our extended model achieved 100% accuracy on the evaluation set, this perfect performance signals a critical overfitting problem rather than a breakthrough in idiomaticity detection. We analyze the key issues below:

Overfitting Analysis The perfect accuracy can be attributed to:

- Small dataset size (70 instances) relative to model complexity
- Over-parameterized architecture [5 $\rightarrow$ 128 $\rightarrow$ 64 $\rightarrow$ 5] for ranking
- Possible learning of dataset-specific patterns rather than generalizable features

Proposed Solutions To address these limitations, we recommend:

- Data augmentation through paraphrasing and context variation
- Simplification of model architecture to reduce parameters
- Implementation of cross-validation for more reliable evaluation

Given these findings, future work should focus on building a more robust evaluation framework and expanding the dataset to ensure real-world generalizability.

5 Conclusion

In this work, we addressed the challenge of idiomaticity detection and representation as part of SemEval 2025 Task 1 (AdMIRe), focusing on text-based approaches. Our system progressed from a baseline implementation using Triplet loss to an enhanced model incorporating cosine similarity computation and a multi-layer neural network architecture.

While our extended model achieved perfect accuracy on the test set, this highlighted important challenges in idiomaticity detection research - particularly the need for larger, more diverse datasets and robust evaluation frameworks. The small dataset size (70 instances) combined with our complex model architecture likely led to overfitting, suggesting that simpler approaches might be more appropriate for the current data constraints.

Several promising directions emerge from this work. First, the development of larger datasets specifically designed for idiomaticity detection could help models learn more generalizable features. Second, exploring data augmentation techniques could help address the limited data availability. Finally, investigating simpler architectures that can effectively balance performance and generalization would be valuable for real-world applications.

Our experience with this task demonstrates both the potential and current limitations of neural approaches to idiomaticity detection. While perfect accuracy on our test set suggests the model can learn to distinguish idiomatic usage in controlled settings, achieving robust, generalizable performance remains an open challenge for future research.

References

[1] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[2] Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I Spy a Metaphor: Large Language Models and Diffusion Models Co-Create Visual Metaphors. In Findings of the Association for Computational Linguistics: ACL 2023, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

[3] Qihao Yang and Xuelin Wang. 2024. FigCLIP: A Generative Multimodal Model with Bidirectional Cross-attention for Understanding Figurative Language via Visual Entailment. In Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), pages 92–98, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

[4] Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024. Enhancing Idiomatic Representation in Multiple Languages via an Adaptive Contrastive Triplet Loss. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.