SemEval-2025 Task 1
AdMIRe: Advancing Multimodal Idiomaticity Representation

Wei He[1], Thomas Pickard[1], Maggie Mi[1], Dylan Phelps[1], Carolina Scarton[1], Marco Idiart[2] and Aline Villavicencio[1,3]

[1] University of Sheffield, UK
[2] Federal University of Rio Grande do Sul, Brazil
[3] University of Exeter, UK

{w.he, tmrpickard1, zmi1, drsphelps1, c.scarton}@sheffield.ac.uk
marco.idiart@gmail.com, a.villavicencio@exeter.ac.uk

Abstract

Effective and accurate representation of non-compositional language is crucial to avoid interpretation errors being propagated to downstream tasks. To evaluate to what extent recent advances in language modelling have improved their ability to identify and interpret non-compositional language and to encourage advances in this area, this task presents the challenge of idiomaticity representation using multimodal data. This task consists of the following subtasks: (A) identifying which of several images best represents an idiomatic expression as it is used in a given sentence, and (B) selecting the best completion for a 3-image sequence representing the meaning of a given expression.

The data consists of text sentences involving idiomatic expressions and images depicting these expressions. This is a follow-up to SemEval-2022 Task 2 which focused on text, but with substantial advances in foundational language models, it is time for more challenging tasks that target semantic understanding in multiple modalities; in this case, static and temporal visual depictions.

1 Overview

Idioms are a class of multi-word expression (MWE) which pose a challenge for current state-of-the-art models because their meanings are often not predictable from the individual words that compose them (Dankers et al., 2022; Villavicencio et al., 2005). For instance, "eager beaver" is unlikely to refer to a passionate muskrat; rather, it typically describes a person who is keen and enthusiastic. These expressions may also generate ambiguity between the literal, surface meaning arising from their component words and the idiomatic meaning. These, among other characteristics, make them a valuable testing ground for examining how NLP models capture meaning.

Motivation Comparing the performance of language models (including large LLMs) to humans shows that models lag behind humans in comprehension of idioms (Tayyar Madabushi et al., 2021; Chakrabarty et al., 2022a; Phelps et al., 2024).

As idioms are believed to be conceptual products and humans understand their meaning from interactions with the real world involving multiple senses (Lakoff and Johnson, 1980; Benczes, 2002), we build on the previous SemEval-2022 Task 2 (Madabushi et al., 2022) and seek to explore the comprehension ability of multimodal models. In particular, we focus on models that incorporate visual and textual information to test how well they can capture representations and whether multiple modalities can improve these representations.

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. For example, due to poor automatic translation of an idiom, the Israeli PM appeared to call the winner of Eurovision 2018 a 'real cow' instead of a 'real darling'! [1]. Our hope is that this task will help the NLP community to better understand the limitations of contemporary language models and to make advances in idiomaticity representation.

2 Task Details

Previous SemEval tasks have explored the evaluation of compositional models (Marelli et al., 2014), paraphrases of noun compounds (Hendrickx et al., 2013) and the interpretation of noun compounds (Butnariu et al., 2009), and more recent tasks have focused on idiomaticity (Madabushi et al., 2022). Other labelled datasets designed for the evaluation of idiomatic and figurative language processing include MAGPIE (Haagsma et al., 2020) and FLUTE (Chakrabarty et al., 2022b).

[1] metro.co.uk

(a) three children in school uniform stealing from a shop

(b) a boy deliberately knocking a cup of tea off a table

(c) a bag of apples

(d) a rotten apple

(e) a sugar-coated peach

Figure 1: Subtask A data example for *bad apple*. Images generated using Midjourney v6.0 (Midjourney, 2024), with a consistent style reference and the prompts shown.

However, as highlighted by Boisson et al. (2023), artifacts present in these datasets may allow models to perform well at the idiomaticity detection task without necessarily developing high-quality representations of the semantics of idiomatic expressions. We present two subtasks which we hope will address these shortcomings by moving away from binary classification and by introducing representations of meaning using visual and visual-temporal modalities.

2.1 Subtask A: Static Images

In Subtask A, participants will be presented with a set of 5 images and a context sentence in which a particular potentially idiomatic nominal compound (NC) appears. The goal is to rank the images according to how well they represent the sense in which the NC is used in the given context sentence.

In order to reduce potential barriers to participation, we also provide a variation of the task in which the images are replaced with text captions describing their content. Two settings are therefore available for the subtask; one in which only the text is available, and one which uses the images.

2.2 Subtask B: Image Sequences (or Next Image Prediction)

Capturing the idiomatic meaning of an MWE in a single image is not necessarily straightforward. While one can envisage a literal *kangaroo court*, a good representation of its idiomatic sense would need to incorporate elements (spontaneity, haste, a potentially predetermined conclusion) which are less concrete than a marsupial wielding a gavel.

In order to better represent the abstract meaning of our target expressions, we generate sequences of 3 images akin to a comic strip, allowing for the depiction of changes in state, mood or relationship between elements over time.

In Subtask B, systems will be given a target expression and an image sequence from which one of the images has been removed, and the objective will be to select the best fill from a sample of images drawn from across our dataset. The NC sense being depicted (idiomatic or literal) will not be given, and this label should also be output.

In order to minimise the risk of non-semantic clues being introduced, the images will adopt a consistent style across the Subtask B dataset. As with Subtask A, we also offer two settings for Subtask B, with descriptive text replacing the images in the 'caption' setting.

3 Data and Resources

Our task uses a potentially idiomatic expression dataset which expands on the SemEval-2022 Task 2 dataset (Tayyar Madabushi et al., 2022), with c. 250 English compounds included. Data are licensed under Creative Commons Attribution-ShareAlike 4.0.

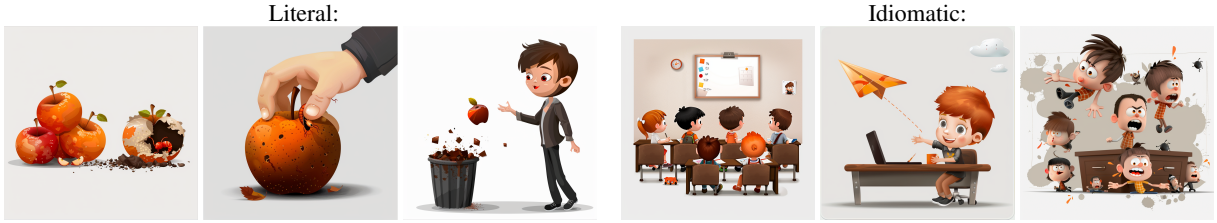
3.1 Subtask A Data

Each idiom generates a set of 5 different images for Subtask A, with a fixed style prompt to encourage consistency. The images for each expression cover a range of idiomaticity:

- A synonym for the idiomatic meaning of the NC.
- A synonym for the literal meaning of the NC.
- Something related to the idiomatic meaning, but not synonymous with it.
- Something related to the literal meaning, but not synonymous with it.
- A 'distractor', which belongs to the same category as the compound (e.g. an object or activity) but is unrelated to both the literal and idiomatic meanings.

Figure 1 shows an example of the Subtask A data for the expression *bad apple*. For a sentence in which *bad apple* is used idiomatically ("The team's efforts were spoiled by the presence of a particular bad apple."), the expectation is that the images will be ordered as shown in Figure 1, with the idiomatic illustration ranked as most similar

Literal: Idiomatic:



(a) A pile of fresh apples on the left. On the right is a rotten apple with a worm coming out of it.

(b) A person picks up the rotten apple with a worm emerging from it.

(c) The person throws the rotten apple into a trashcan.

(d) A group of well-behaved children in a classroom.

(e) A child throwing a paper plane in the classroom.

(f) A group of children behaving badly in a classroom.

Figure 2: Subtask B data example for *bad apple*. Images generated using Midjourney v6.0 (Midjourney, 2024), with consistent style reference prompts shown. These prompts were generated by Gemini Pro 1.5 (Gemini Team, 2023) with adjustments.

to the in-context sense (this is the most important metric for evaluation).

For a literal interpretation, the expected order would be reversed, except for the distractor item (*a sugar-coated peach*), which should remain unrelated to either interpretation.

3.2 Subtask B Data

Two sequences of images are generated for each NC: one sequence representing the literal and one the idiomatic meaning (Figure 2). Each image in a sequence is generated individually using prompts crafted by an instruction-tuned text-to-text generation model (Gemini Pro), inspired by the work of Chakrabarty et al. (2023) on visual metaphors, and styled consistently for uniformity across the data.

An example of the Subtask B data for the expression *bad apple* is shown in Figure 2. The corresponding image captions are shown in Appendix A.

Data Quality and Ethics Data quality will be ensured by measuring agreement between human reviewers of the generated data, with low-quality items filtered out. Context sentences containing target expressions are obtained from web sources or specifically written, and fall within the four factors of fair use: the data is used for non-profit research purposes; publicly available; the amount of text used is a very small fraction of the original piece; and does not impact the marketability of the original content. There are no privacy concerns with respect to the data used as we do not use any data associated with individuals and all annotation is performed with the ethics clearance of the University of Sheffield.

4 Evaluation

Human benchmarks will be obtained for all task configurations.

4.1 Subtask A

Performance for Subtask A will be assessed with two key metrics:

- Top Image Accuracy: Correct identification of the most representative image.
- Rank Correlation: Spearman's rank correlation of model rankings with ground truth.

4.2 Subtask B

This subtask assesses the model's ability to complete a sequence of images that narratively represent an idiomatic expression, along with distinguishing between idiomatic and literal meanings. Evaluation metrics will be:

- Completion Accuracy: Correctly selecting the image to complete the narrative.
- Labeling F1 Score: Effectiveness in identifying idiomatic versus literal expressions.

5 Task organisers

Prof Aline Villavicencio University of Exeter (UEx) and Sheffield (UShef), UK. She is a member of the editorial board of Computational Linguistics, TACL and of JNLE.
a.villavicencio@exeter.ac.uk

Prof Marco Idiart Federal University of Rio Grande do Sul (Brazil). Research interests include MWEs and neural networks. marco.idiart@gmail.com

Dr Carolina Scarton (UShef). Research interests in social media analysis, machine translation and multiword expressions.c.scarton@sheffield.ac.uk

Dr Wei He (UShef). Research interests include computational linguistics and deep learning. w.he@sheffield.ac.uk.

Maggie Mi, **Dylan Phelps** and **Thomas Pickard** (UShef) {zmi1, drsphelps1, t.pickard}@sheffield.ac.uk.

References

Réka Benczes. 2002. The semantics of idioms: a cognitive linguistic approach. *The Even Yearbook*, 5:17–30.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction Artifacts in Metaphor Identification Datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022a. 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.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative Language Understanding through Textual Explanations. ArXiv:2205.12404 [cs].

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.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, pages 279–287, Marseille, France. European Language Resources Association.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. ArXiv:2204.10050 [cs].

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Midjourney. 2024. Midjourney.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377. Special issue on Multiword Expression.

A Complete Training Example

The complete sample training data for the idiomatic expression *bad apple* are shown below. Note that the image generation prompts and sense labels (idiomatic/literal) will not be made available to participants, and image file names will be randomised to prevent data leakage.

A.1 Task A Sample Data

Config:	Idiom image	Config:	Idiom related	Config:	Literal related
Prompt:	three children in school uniform stealing from a shop	Prompt:	a boy deliberately knocking a cup of tea off a table	Prompt:	a bag of apples
Image:		Image:		Image:	
Caption:	The image shows an illustration of two children, a boy and a girl, standing in front of a small store or kiosk. The store has a thatched roof and appears to be selling various items, including what looks like plants and possibly some food items, as suggested by the presence of a microwave and a sandwich. The children are dressed in school uniforms, which include ties, suggesting they might be on their way to or from school. The scene is set outdoors, and the children seem to be engaged in a conversation or transaction with the store.	Caption:	The image shows an animated character, a young boy with spiky hair, who appears to be in a state of surprise or shock. He is standing at a table with a cup of coffee that has been knocked over, causing coffee to spill onto the table and the floor. The boy's expression and the splashing coffee suggest a sudden, unexpected event.	Caption:	The image shows a basket filled with ripe, orange apples. The apples have a glossy finish and are adorned with green leaves. The basket appears to be made of a woven material, possibly burlap, and is placed on a surface with a few fallen apples and leaves scattered around it. The overall scene suggests a harvest or a display of fresh produce.
Config:	Literal image			Config:	Distractor
Prompt:	a rotten apple			Prompt:	a sugar-coated peach
Image:				Image:	
Caption:	The image shows a stylized illustration of an apple with a bite taken out of it. The apple is depicted with a realistic texture, and the bite reveals a brown interior with a few seeds visible. The apple is also shown with a green leaf attached to its stem, which is still attached to the apple. The background is plain white, which highlights the apple and its details.			Caption:	The image shows a highly stylized and artistic representation of a peach. It features a peach with a vibrant orange color, a green leaf attached to the top, and a brown stem. The peach is cut open to reveal its juicy interior, which is also depicted in a realistic manner. The background is a plain white, which contrasts with the peach and highlights its details. The image has a smooth, almost glossy texture, and the lighting gives it a soft, almost ethereal quality.

Table 1: Subtask A data sample for *bad apple*.

A.2 Task B Sample Data

		Literal			
Prompt:	A pile of fresh apples on the left. on the right is a rotten apple with a worm coming out of it.	Prompt:	A person picks up the rotten apple with a worm emerging from it.	Prompt:	The person throws the rotten apple into a trashcan.
Image:		Image:		Image:	
Caption:	cartoon vector illustration of an apple, orange and chocolate in the shape of apples with one half broken open to reveal crumble inside, white background, side view	Caption:	hand pulling out an earthworm from the core of rotten apple	Caption:	A cartoon vector illustration of an adult man with brown hair and a short beard wearing business throws away an apple into a garbage bin full of chocolate pieces isolated on a white background.
		Idiomatic			
Prompt:	A group of well-behaved children in a classroom	Prompt:	A child throwing a paper plane in the classroom	Prompt:	A group of children behaving badly in a classroom
Image:		Image:		Image:	
Caption:	A group of children sitting at desks in the classroom, whiteboard on wall	Caption:	A cute little boy sits at his desk, playing with paper airplane toys in his hands while using his laptop computer. The background is simple and clean with a bright, transparent texture style. The illustration is in the style of a cartoon character with vector graphics. White clouds float in the sky above his head in a cartoon style. The design is a cute cartoon in high resolution on a simple, pure gray background with an isometric view and bright colors.	Caption:	cartoon character design sheet of angry kids jumping on desk, different poses and expressions, cute style

Table 2: Subtask B data sample for *bad apple*.