ChuenSumi at SemEval-2025 Task 1: Sentence Transformer Models and Processing Idiomacity

Sumiko Teng
Waseda University
sumiko@fuji.waseda.jp

Chuen Shin Yong
Waseda University
yongchuenshin@suou.waseda.jp

Abstract

This paper participates Task 1 of SemEval2025, specifically Subtask A's English Text-Only track, where we develop a model to rank text descriptions of images with respect to how well it represents a the use of a given multi-word expression in its respective context sentence. We trained sentence transformer models from huggingface to rank the text descriptions, finding the RoBERTa model to be the better performing model. For the final evaluation, the fine-tuned RoBERTa model achieved an accuracy of 0.4 for the first developer's evaluation set, and 0.2 for the second, ranking 9th in the English Text Only category for Subtask A. Overall, our results show that a vanilla sentence transformer approach performs adequately in the task and processing idioms. They also suggest that RoBERTa models may be stronger in idiom processing than other models.

1 Introduction

Multiword expressions (MWEs), such as idioms, are prevalent in natural language. They occur frequently in all domains (Biber et al., 2021) and constitute a significant portion of any speaker's lexicon, comparable to portions of single-word expressions (Jackendoff, 1997). Thus, it is important that language models can effectively process idiomatic MWEs. However, studies show that computational models struggle with idiom comprehension, especially when compared to human performance (Phelps et al., 2024; Tayyar Madabushi et al., 2021). This difficulty arises because the meaning of idioms often cannot be predicted based on the combination of the meanings of their individual parts (Dankers et al., 2022). Thus, Task 1 of SemEval2025 focuses on improving current models of idiom comprehension.

Specifically, we participate in Subtask A's English Text-Only track, where we are required to develop a model which ranks text descriptions of images with respect to how well it represents a given MWE in its respective context sentence. To complete the task, we fine-tuned two sentence transformer models from huggingface to take the sentence with the given MWE and its respective text descriptions as inputs, then produce the rankings of the text descriptions as outputs. One model was an mpnet model, while the other was a RoBERTa model. We found that the RoBERTa model produced higher top image accuracy and Spearman's Rank Correlation scores.

During the development phase, we also experimented with a split approach. This approach consisted of first training a standard BERT model to work as a binary classifier to classify an MWE as idiomatic or literal based on its use in its context sentence. Then, text descriptions are scored based on their idiomacity levels using ranking boosting algorithm. Finally, the text descriptions were ranked based on their scores. However, this approach did not yield significant results, hence is not elaborated on in detail in this paper.

The trained RoBERTa model was the model used for the final evaluation. It achieved an accuracy of 0.4 for the first developer's evaluation set, and 0.2 for the second, ranking 9th in the English Text Only category for Subtask A. Overall, our results show that a vanilla sentence transformer approach performs adequately, but further optimisations can be explored to enhance performance. Our code is available on GitHub[1].

2 Background

In this section, we give an overview of the relevant literature, subtask and dataset.

[1] https://github.com/svmiko/semeval25-task1/tree/874fa5f04d823d3e3f22b41023bc216f75d1ce2e/system

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 122–126
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2.1 Relevant Literature

Some studies on idiom processing attempted to improve model comprehension by encouraging a more compositional analysis (Li et al., 2021; Raunak et al., 2019). However, later studies suggest that such compositional approaches reduce idiom comprehension.

Hence, more recent approaches encourage leveraging an idiom's surrounding context or treating idioms as single lexical units instead. These approaches align with linguistic research, which suggests that humans tend to process idioms as a single unit rather than as compositional sequences (Sinclair, 1991). Lakoff and Johnson (1980) also shows how idioms derive meaning from their context and real-world interactions. Thus, newer studies have used masked language modelling, which encode richer contextual information, demonstrating improvements in handling non-compositional semantics such as idioms (Fakharian and Cook, 2021; Zeng and Bhat, 2021). Many studies have also shown that encoding idioms as single entities result in better processing of them (Chakrabarty et al., 2023; Tayyar Madabushi et al., 2021; Zaninello and Birch, 2020). Building on these findings, Tayyar Madabushi et al. (2022) fine-tuned a sentence transformer model incorporating single-token representations of idioms, achieving strong performance in idiom comprehension tasks. These findings suggest that utilising an idiom's textual context is a promising direction for improving language model performance in idiom processing.

This task helps to build on existing work to improve machines' comprehension of idioms.

2.2 Task and Dataset

Subtask A is essentially a ranking task. Participants are given a context sentence containing a potentially idiomatic nominal compound (NC), alongside 5 images and respective text descriptions of the images. The objective is to rank the images or their text descriptions based on how well they capture the meaning of the NC in the given context. For this paper, we participate using only the text descriptions, without the images. Participants were ranked based on two criteria. First, top image accuracy, which refers to how accurately the developed system identifies the most representative image, or text description for the context sentence. Second, based on Spearman's rank correlation of the ranks generated by the model and those by the developers.

Depending on whether the target NC is used idiomatically or literally, the developer's desired ranking changes accordingly. Before discussing the rankings, we first describe how the images' respective text descriptions are related to the NC. As there are five images per NC, there are also five corresponding text descriptions. They can describe: (1) an idiomatic synonymous use of the NC, (2) an idiomatic non-synonymous use of the NC, (3) literal synonymous use of the NC, (4) a literal non-synonymous use of the NC, or (5) be unrelated to the NC. If the NC is used *idiomatically*, the developers rank the images and text descriptions as follows:

1. Highest ranked, most representative of use of NC in context sentence: The image and text description that depicts the NC in an *idiomatic* and *synonymous* manner.

2. Image and text description that depicts the NC in an *idiomatic* and non-synonymous manner.

3. Image and text description that depicts the NC in an literal and *synonymous* manner.

4. Image and text description that depicts the NC in an literal and non-synonymous manner.

5. Lowest ranked, least representative use of NC in context sentence: Image and text description unrelated to NC.

Conversely, when the NC is used literally:

1. Highest ranked: Image and text description that depicts the NC in an literal and *synonymous* manner.

2. Image and text description that depicts the NC in an literal and non-synonymous manner.

3. The image and text description that depicts the NC in an *idiomatic* and *synonymous* manner.

4. Image and text description that depicts the NC in an *idiomatic* and non-synonymous manner.

5. Lowest ranked: Image and text description unrelated to NC.

Text descriptions unrelated to the NC serve as a distractor, hence are always ranked least similar to the context sentence by the developers (Pickard et al., 2025). To summarise, these rankings are the rankings provided in the developers' training dataset.

The dataset used for this paper is the English Subtask A training dataset with text descriptions provided by the developers. It contains 70 NCs and their respective context sentences. Out of these 70 NCs, 39 are used idiomatically, while 31 are used literally, making it a small, but relatively balanced dataset. As each NC has 5 respective text descriptions, there were 350 text descriptions in total. While no information on how the text descriptions were generated was provided (i.e. we do not know if these text descriptions were written by the developers themselves or generated using AI models), they seem to have been written following a similar style.

3 System Overview

Training datasets provided by the developers were used to fine-tune the selected models. More details on the selected models will be given in Section 4 “Experimental Set-up.” The data we used were context sentences, nominal compounds (NC), text description of related images, and the rankings of text descriptions in terms of how well they represent the use of the NC in the context sentence. When pre-processing our dataset, we labeled the text descriptions as "candidates," and the NCs as "compound." Other than these labels, no preprocessing was conducted on the text descriptions and context sentences, as the models selected for fine-tuning were sentence transformer models, which have been shown to handle raw textual input effectively (Agirre et al., 2016). See Table 1 for an example of our dataset.

Sentence transformer models were fine-tuned to perform ranking for the subtask. We chose to use sentence transformer models as they have been shown to perform well in ranking tasks (Di Liello et al., 2022). Both the candidates and context sentences were used as input, so the model could directly learn the relationship between the context sentence and candidates. They were also grouped by their respective compounds to ensure that candidates are ranked only in relation to their respective NC-containing context sentence. Embeddings for context sentences and candidates are generated using each model's respective encoder, and cosine similarity is used to measure the semantic proximity between the candidates and context sentences. Based on these similarity scores, candidates are ranked from 1 to 5, with 1 being the most similar and 5 being the least. These ranks serve as the model's output during inference. The loss function used when fine-tuning is CosineSimilarityLoss, which optimises similarity-based ranking, making it useful for a ranking task (see Reimers and Gurevych, 2019).

While sentence transformer models provide a robust framework for ranking candidates based on semantic similarity, one key challenge of the task was semantic ambiguity. Idiomatic expressions often exhibit semantic ambiguity, meaning that the same phrase can be interpreted differently based on the surrounding context. Hence, our system leverages contextual embeddings generated from sentence transformer models that capture the nuanced relationship between the context sentence and each candidate. The model does not treat candidates in isolation but instead encodes their meaning in relation to the context sentence. Additionally, fine-tuning with CosineSimilarityLoss ensures that candidates are ranked based on their semantic proximity to the context, allowing the model to learn fine-grained distinctions between literal and idiomatic uses.

A limited dataset presents challenges in ensuring good model performance, especially when the test data can contains nominal compounds that were not seen during training. We designed the system to learn from the relationship between context and candidate sentences rather than memorizing specific nominal compounds. By focusing on general patterns of semantic similarity across all instances, the model is better equipped to generalize to unseen nominal compounds.

4 Experimental Setup

For this task, we decided to work on getting embeddings to understand how candidate sentences may have literal and idiomatic representations. We fine-tuned two pre-trained sentence transformer models to generate embeddings for the ranking task: "paraphrase-multilingual-mpnet-base-v2," henceforth the "mpnet model", and "sentence-transformers/all-roberta-large-v1," henceforth the "RoBERTa model." Both models were taken from Huggingface. The dataset was prepared by splitting the available data using the train-test-split function, following an 80-20 split, where 80 percent of the data was allocated for training and 20 percent for testing. As the data was grouped by compounds

Compound	Context_Sentence	Candidate	Ranking
night owl	...I am a night owl, so I find that going to sleep...	The image depicts a nighttime scene...	4
night owl	...I am a night owl, so I find that going to sleep...	The image depicts a cartoon-style illustration of a person...	1
night owl	...I am a night owl, so I find that going to sleep...	The image depicts a cartoon-style owl...	3
night owl	...I am a night owl, so I find that going to sleep...	The image depicts a cartoon-style illustration of a young...	2
night owl	...I am a night owl, so I find that going to sleep...	The image depicts a dumbbell...	5

Table 1: Example of Dataset

(see section 3), the training set consisted of 56 compounds and their respective text descriptions and context sentences, while the test set contained 14 compounds grouped in a similar manner. As previously mentioned, the CosineSimilarityLoss function was used to optimise ranking performance. For the mpnet model, it was initially set to fine-tune for 30 epochs, but the process was terminated early to prevent overfitting. 100 warm-up steps were also applied for stable optimisation. Based on the results of training the mpnet model, the RoBERTa model was fine-tuned for only 10 epochs. 100 warm-up steps and 500 evaluation steps were also incorporated. The AdamW optimizer was used to enhance weight regularisation and prevent gradient-based overfitting. To evaluate the fine-tuned models on the test set, we calculated Top Image Accuracy and Spearman's Rank Correlation, as required by the task (see section 2.2).

5 Results

The RoBERTa model performed better than the mpnet model in the ranking task (see table 2).

Table 2: Comparison of Model Performance

Model	Top-1 Accuracy	Spearman's ρ
MPNet (multilingual)	50.00%	0.4643
RoBERTa	57.14%	0.5071

Based on evaluation on the test set, the RoBERTa model had 57.1 percent Top Image Accuracy, and a Spearman's Rank Correlation of 0.507. Conversely, the mpnet model had a Top Image Accuracy of 50 percent and a Spearman's Rank Correlation of 0.464. Therefore, we retrained the full training dataset provided and submitted the results of the RoBERTa model for the final evaluation. For the final evaluation, we achieved an accuracy of 0.4 for the first developer's evaluation set, and 0.2 for the second, ranking 9th in the English Text Only category for Subtask A.

6 Conclusion

Our results show that a vanilla sentence transformer approach performs adequately, but further optimizations can be explored to enhance performance. We initially experimented with a split approach and more complex systems, which are:

1. Training a binary classifier to determine whether a context sentence is idiomatic or literal (using standard BERT).
2. Scoring candidates based on their idiomaticity level using ranking boosting algorithms.
3. Ranking candidates based on their scores or experimenting with Siamese networks with a custom loss function for rankings.

However, this approach did not yield significant improvements over the direct ranking method. Future work could explore hybrid architectures that combine classification-based pre-filtering with ranking models, as well as larger pre-trained models trained on more extensive idiomatic datasets. Additional customisation in loss functions, feature engineering, and ensemble methods may also improve ranking accuracy.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016*.

10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics).

Douglas Biber, Stig Johansson, Geoffrey N Leech, Susan Conrad, and Edward Finegan. 2021. *Grammar of spoken and written English*. John Benjamins.

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. *arXiv preprint arXiv:2305.14724*.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Pre-training transformer models with sentence-level objectives for answer sentence selection. *arXiv preprint arXiv:2205.10455*.

Samin Fakharian and Paul Cook. 2021. Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th workshop on multiword expressions (mwe 2021)*, pages 23–32.

Ray Jackendoff. 1997. Twistin'the night away. *Language*, pages 534–559.

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

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. *arXiv preprint arXiv:2105.14802*.

Dylan Phelps, Thomas Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. *arXiv preprint arXiv:2405.09279*.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, Carolina Scarton, and Marco Idiart. 2025. Semeval-2025 task 1: Admire - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluations (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Vikas Raunak, Vaibhav Kumar, and Florian Metze. 2019. On compositionality in neural machine translation. *arXiv preprint arXiv:1911.01497*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.

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 preprint arXiv:2204.10050*.

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. *arXiv preprint arXiv:2109.04413*.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3816–3825.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.