ANLP PROJECT REPORT

Amey Choudhary, Aryan Gupta, Sarthak Bansal

November 20, 2024

1 Introduction

Effective representation of idioms is crucial for applications such as sentiment analysis, machine translation, and natural language understanding. Enhancing models' ability to interpret idiomatic expressions can significantly improve the performance of these applications. For instance, poor automatic translation of an idiom once caused the Israeli Prime Minister to describe the winner of Eurovision 2018 as a "real cow" instead of a "real darling."

When comparing the performance of language models (including large LLMs) with humans, models consistently lag in comprehending idioms [TMGSSV21]. Idioms are conceptual constructs requiring interactions between entities in a sentence and real-world knowledge to grasp their meaning. This project focuses on models that process textual information to interpret idiomatic expressions effectively. We evaluate the comprehension abilities of these models and explore approaches to enhance their performance.

Specifically, the task involves ranking five different captions based on their similarity to a query sentence. The sentences include compound words—idioms—making the task particularly focused on developing models that accurately interpret idiomatic usage which can be used literally or idiomatically. For e.g. query sentence "Kapil ate a rotten apple and felt disgusted" contains literal meaning of the compound word ("rotten apple") while the query sentence "The students in this class are good except for a few rotten apples" uses the compound word ("rotten apples") in an idiomatic sense.

To this end, we propose a variety of models trained on the dataset and evaluate them using three distinct metrics. In the subsequent sections, we discuss recent advancements in understanding idiomatic usage, examine the characteristics of the dataset, detail our implementation, and present the results of our experiments.

2 Literature Survey Overview

Representing idiomaticity in language models remains a complex challenge, particularly given the non-compositional nature of idiomatic expressions. Idioms often convey figurative meanings that cannot be directly inferred from their constituent words, making their representation distinct from traditional compositional phrases ([KC18];[CVIR19];[MGSG+22]). While large language models like GPT-3 achieve substantial success in general NLP tasks, their performance in idiom comprehension remains subpar, achieving only 50.7% accuracy in understanding idiomatic expressions [ZB22].

Previous efforts to address idiomaticity have explored various strategies, such as combining compositional components with adaptive weights ([HT16];[LYW+18]), creating phrase embeddings that adaptively capture both compositional and idiomatic meanings ([HT16]), or using idiom-specific tokens for representation ([YS15]; [LLXL18]). More recently, adapter-based approaches have been proposed to augment pre-trained models like BART, equipping them with the ability to better capture idiomatic meanings ([ZB22]). Methods such as PIER incorporate "idiomatic adapters" that specialize in learning representations for non-compositional phrases, significantly improving performance in idiom-related tasks ([ZB23]).

Idiomatic representation has been assessed using both intrinsic and extrinsic evaluation methods. Intrinsic evaluations focus on the internal consistency of idiomatic representations, using datasets such as AStitchInLanguageModels ([TMGSSV21]) and the Noun Compound Type and Token Idiomaticity dataset ([GKVS+21]) to measure how well models capture idiomaticity. These evaluations often rely on semantic similarity tasks and probing techniques to test how closely model-generated embeddings align

with human judgments. Extrinsic evaluations, on the other hand, examine the downstream impact of idiomatic representations on various tasks, including machine translation and sentiment analysis ([DLT22]), sentence generation ([ZGB21]), conversational systems([ALL22]).

A standard evaluation methodology is idiomaticity classification at the sentence or token level. Earlier studies have approached this through various techniques ([HBN20];[SP20]). These methods typically involve determining whether an expression is used idiomatically or literally in context.

More comprehensive tasks, such as SemEval-2022 Task 2B ([MGSG+22]), extend the evaluation to a multilingual setting. This task evaluates idiomaticity representation by requiring models to predict Semantic Textual Similarity (STS) scores between sentence pairs, whether or not they contain idiomatic expressions. Such tasks emphasize not only accurate idiomatic representation but also the ability of models to generalize across languages and contexts. The use of STS ensures that the models are evaluated on their semantic understanding rather than mere memorization of idioms.

Recent advancements also explore specialized metrics for non-compositional expressions, such as idiomaticity-specific loss functions and contrastive learning frameworks ([ZZB23]), which optimize models for distinguishing between literal and idiomatic meanings in text. These innovations further improve the precision of both intrinsic and extrinsic evaluations, enhancing the robustness of idiomatic representation across diverse NLP tasks.

3 Dataset

The dataset consists of 70 samples each of which contains 5 caption sentences and 1 query sentence along with the sentence type (i.e. usage of idiom - literal or idiomatic) and the ground truth ranking of captions w.r.t query sentence. Here are some observations from the dataset :

- **Word Frequency and Distribution** : Idiomatic sentences contain a greater number of unique words than literal sentences. This can be attributed to a richer vocabulary in sentences with idiomatic usage indicating that descriptions for idiomatic imagery are generally more varied and potentially more descriptive or imaginative.
- **Sentence Length**: The length of sentences for both idiomatic and literal sentences tends to be almost the same. This indicates that the complexity or depth of content does not significantly impact sentence length in either category.
- **Common Words**: In idiomatic sentences common words such as "was," "no," "your," and "them" imply a narrative or explanatory style in sentences. Conversely, words like "up," "his," and "this" found in literal sentences suggest more direct or present descriptions.

4 Methodology

4.1 Baseline approach

We observe that the idiomatic usage of compound words varies with the context in which they are used. Consequently, it is intuitive to employ an encoder that generates context-aware pre-trained embeddings.

As a baseline approach, we utilize a pre-trained Sentence-BERT model functioning as a dual encoder to create embeddings for both the query sentence and the five captions. The cosine similarity between these dense representations is then used to rank the captions.

4.2 Objective Function

To enhance the embeddings of the pre-trained model for better interpretation of *idiomatic usage*, it is essential to align the model with our specific dataset. This requires the design of a convex objective function that incorporates both the ground truth ranking of captions and the embeddings generated by the model. The primary goal is to clearly distinguish higher-ranked captions from lower-ranked ones.

Specifically, we aim to ensure that the embeddings are sufficiently distinct to enable accurate ranking without ambiguity. Furthermore, the difference in rank should influence the distance between

embeddings. For instance, the embedding of a caption ranked 2nd should be closer to the embedding of the caption ranked 1st than to the embedding of the caption ranked 5th.

To achieve this, we propose a modification to the adaptive triplet loss, which we call the **Ranked Triplet Loss**. This loss function is designed to leverage the relative ranking of captions by emphasizing the differences between higher-ranked and lower-ranked captions in terms of their embedding distances from the query sentence embedding. Additionally, instead of using the embedding of the entire sentence, we extract embeddings specifically from the position where the compound word (idiomatic expression) appears.

The **Ranked Triplet Loss** is implemented as follows: Given the query sentence embedding and the embeddings of multiple captions (at the position where the idiomatic compound is located), we iterate through all possible pairs of captions. For each pair, we compute the similarity between the query embedding and the caption embeddings using cosine similarity. Let d_{pos} represent the similarity between the query and a higher-ranked caption, and d_{neg} represent the similarity with a lower-ranked caption. The triplet loss for each pair is computed as:

$$\text{triplet\_loss} = \frac{1}{|i-j|} \cdot \text{ReLU}(d_{\text{neg}} - d_{\text{pos}} + \text{margin}),$$

where i and j are the positions of the captions in the ranking, and the weight $\frac{1}{|i-j|}$ ensures that the loss is inversely proportional to the difference in their rankings. This weighting mechanism gives more importance to captions that are closer in rank, thus aligning their embeddings more tightly. The total loss is then computed as the mean of all pairwise triplet losses across the dataset.

4.3 Data Preprocessing

As mentioned earlier, we aim for the embeddings to focus on the position where the idiom is used rather than the entire sentence. To enforce this within the loss function, we additionally *highlight* the idioms by placing [SEP] tokens around the idiomatic compound. This explicit marking allows the model to differentiate the idiom from the rest of the sentence, thereby improving its ability to interpret the idiomatic usage in context.

4.4 Fine-tuning Approaches

Literature suggests that BERT is a contextual model that produces excellent results when fine-tuned for specific tasks. However, due to its large size, fine-tuning BERT requires significant computational and memory resources. To address this challenge, alternative techniques have been proposed to fine-tune a smaller subset of parameters while still effectively encapsulating the task objective. These techniques aim to retain BERT's contextual understanding while reducing the overhead associated with full model fine-tuning.

4.4.1 LoRA

Low-Rank Adaptation (LoRA) is used to efficiently fine-tune the BERT model without updating all its parameters. LoRA introduces trainable low-rank matrices into the transformer layers, enabling task-specific adaptations with minimal computational overhead. This approach is particularly effective for adapting large pre-trained models like BERT to specific tasks such as understanding idiomatic expressions, where full fine-tuning may be impractical. Additionally, by capturing *attention updates*, we expect LoRA to effectively capture the contextual information necessary to interpret idiomatic usage.

4.4.2 CNN

Convolutional Neural Networks (CNNs) are employed to capture local dependencies and n-gram patterns within the embeddings generated by BERT. By applying convolutional filters to the embeddings, CNNs help identify features that are critical for interpreting idiomatic expressions. These features often depend on subtle local patterns within the text, making CNNs a valuable addition to the fine-tuning process.

4.4.3 BiLSTM

Bidirectional Long Short-Term Memory (BiLSTM) networks are incorporated to model sequential and contextual dependencies in the text. Unlike unidirectional models, BiLSTM processes the input in both forward and backward directions, capturing information from the entire sentence. This bidirectional context is crucial for understanding idiomatic usage, as the meaning of idioms often depends on the surrounding words and their relationships.

5 Results

5.1 Evaluation Metrics

5.1.1 Spearman Score

The *Spearman rank correlation coefficient*, denoted as ρ, is a non-parametric measure of the strength and direction of association between two ranked variables. It evaluates how well the relationship between the variables can be described using a monotonic function.

The formula for Spearman's rank correlation is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where:

- n is the number of paired ranks,
- d_i is the difference between the ranks of the i-th observation in the two variables,
- $\sum d_i^2$ is the sum of the squared rank differences.

The correlation ranges from $\rho = -1$ (perfect negative correlation) to $\rho = 1$ (perfect positive correlation), with $\rho = 0$ indicating no monotonic relationship. It is widely used in non-parametric statistics and ordinal data analysis.

5.1.2 Kendall Score

The *Kendall rank correlation coefficient*, denoted as τ, is a non-parametric measure of the strength and direction of association between two variables. It evaluates the similarity of the orderings of the data.

The formula for Kendall's τ is:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}$$

where:

- C is the number of concordant pairs (pairs where the rank order is the same for both variables),
- D is the number of discordant pairs (pairs where the rank order is different),
- T_x and T_y are the number of ties in the first and second variable, respectively.

The value of τ ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with $\tau = 0$ indicating no correlation.

5.1.3 Comparative Analysis of Metrics

Each of the metrics—Spearman's rank correlation (ρ) and Kendall's rank correlation (τ)—differs in its sensitivity to the positions of items within a ranking. Both metrics are based on the relative ordering of items, but they differ in their emphasis and interpretation.

Spearman's ρ evaluates the global rank order of items by considering all pairs in the list and assessing how well the ranks from two different lists align overall. It measures the monotonic relationship

Model Name	Validation Loss	Spearman Score	Kendall Score	Trainable Parameters
Baseline	-	0.116	0.083	-
BERT(full finetuned)	**0.462**	**0.537**	0.426	109M
BERT(frozen)+LoRA	0.568	0.291	0.229	147K
BERT(frozen)+CNN	0.660	0.518	**0.434**	590K
BERT(frozen)+CNN+LoRA	0.609	0.258	0.271	884K
BERT(frozen)+BiLSTM	0.579	-0.233	-0.159	526K
BERT(frozen)+BiLSTM+LoRA	0.552	0.035	0.044	821K
xlm_RoBERTa(full finetuned)	0.658	-0.189	-0.129	277M
xlm_RoBERTa(frozen)+LoRA	0.632	0.291	0.231	147K

Table 1: Model Performance Metrics : Best results are highlighted in **bold**

between two rankings, focusing on whether the overall order between items is preserved. Spearman's ρ is less sensitive to small changes in the ranks of items that are close to each other and instead emphasizes the consistency of the ranking as a whole.

In contrast, Kendall's τ focuses on local rank relationships by counting the number of concordant and discordant pairs of items. A pair is concordant if the relative order of the items is the same in both rankings, and discordant if it differs. Kendall's τ places more importance on local changes in rank and is more sensitive to discrepancies in the ordering of adjacent or nearby items.

Thus, while Spearman's ρ prioritizes global consistency in rankings, Kendall's τ emphasizes local rank relationships, providing a finer-grained measure of rank agreement.

5.2 Quantitative Results

We investigated the performance of various transformer-based models for ranking sentences based on their idiomatic usage relative to an anchor sentence. All models were trained on a dedicated training dataset for 200 epochs, and the scores reported were evaluated on the validation dataset. The results are summarized in Table 1.

The **fully fine-tuned BERT model** exhibited the best performance, indicating that BERT's contextual embeddings are highly effective in capturing idiomatic interpretability. However, the significant computational and memory overhead associated with fine-tuning all 109 million parameters makes it impractical for real-world applications where efficiency is a priority.

The **BERT (Frozen) + CNN** configuration demonstrated competitive performance, outperforming most other fine-tuning approaches while maintaining a much smaller number of trainable parameters. This highlights the ability of convolutional layers to efficiently capture local contextual interactions within sentences, making it a suitable and lightweight alternative to full fine-tuning.

In contrast, the **BERT (Frozen) + BiLSTM** configuration underperformed, even scoring lower than the baseline. This result can be attributed to the sequential modeling of BiLSTM introducing unnecessary noise into the task. Since idiomatic usage often relies on precise contextual information rather than long-range dependencies, the sequential nature of BiLSTM may have misaligned with the requirements of the task, leading to confusion in rankings.

Surprisingly, **LoRA** did not significantly improve the performance of the base model in most configurations, including BERT and CNN. However, it showed some improvement when paired with BiLSTM and XLM-RoBERTa, though these results were still relatively poor compared to other approaches. This suggests that updating attention mechanisms alone may play a limited role in capturing idiomatic information. While LoRA does not produce entirely negative results, its impact on creating agreeable rankings is minimal, indicating that idiomatic interpretation relies more heavily on the base embeddings and local contextual patterns than on fine-grained attention updates.

In summary, the fully fine-tuned BERT model remains the most effective approach but is computationally expensive. The **BERT (Frozen) + CNN** configuration offers a strong balance between performance and efficiency, making it the most practical alternative for real-world applications. Sequential models like BiLSTM and lightweight fine-tuning approaches such as LoRA may require further optimization or task-specific adaptations to better handle idiomatic ranking tasks.

6 Conclusion

In this work, we focus on developing models that can effectively capture idiomatic usage in a sentence. To this extent, we designed a unique loss function, tried simple data augmentation, and experimented with various fine-tuning techniques found in the literature. We observe that capturing contextual dependencies is important for this task, which BERT is able to achieve effectively when fully fine-tuned. However, lighter alternatives like CNNs exist which appear as strong practical contenders.

References

[ALL22] Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. Vector representations of idioms in conversational systems. *Sci*, 4(4):37, 2022.

[CVIR19] Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57, 03 2019.

[DLT22] Verna Dankers, Christopher Lucas, and Ivan Titov. Can transformer be too compositional? analysing idiom processing in neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[GKVS+21] Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online, August 2021. Association for Computational Linguistics.

[HBN20] Hessel Haagsma, Johan Bos, and Malvina Nissim. Magpie: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France, 2020. European Language Resources Association.

[HT16] Kazuma Hashimoto and Yoshimasa Tsuruoka. Adaptive joint learning of compositional and non-compositional phrase embeddings. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany, August 2016. Association for Computational Linguistics.

[KC18] Milton King and Paul Cook. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[LLXL18] Minglei Li, Qin Lu, Dan Xiong, and Yunfei Long. Phrase embedding learning based on external and internal context with compositionality constraint. *Knowledge-Based Systems*, 152:107–116, 2018.

[LYW+18] Bing Li, Xiaochun Yang, Bin Wang, Wei Wang, Wei Cui, and Xianchao Zhang. An adaptive hierarchical compositional model for phrase embedding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4144–4151, 2018.

[MGSG+22] Harish Tayyar Madabushi, Edward Gow-Smith, Marcos García, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 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, 2022.

[SP20] Prateek Saxena and Soma Paul. Epie dataset: A corpus for possible idiomatic expressions. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors, *Text, Speech, and Dialogue*, pages 87–94, Cham, 2020. Springer International Publishing.

[TMGSSV21] Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[YS15] Wenpeng Yin and Hinrich Schütze. Discriminative phrase embedding for paraphrase identification. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1368–1373, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[ZB22] Ziheng Zeng and Suma Bhat. Getting bart to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137, 2022.

[ZB23] Ziheng Zeng and Suma Bhat. Unified representation for non-compositional and compositional expressions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11696–11710, Singapore, December 2023. Association for Computational Linguistics.

[ZGB21] Jianing Zhou, Hongyu Gong, and Suma Bhat. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In Paul Cook, Jelena Mitrović, Carla Parra Escartín, Ashwini Vaidya, Petya Osenova, Shiva Taslimipoor, and Carlos Ramisch, editors, *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online, August 2021. Association for Computational Linguistics.

[ZZB23] Jianing Zhou, Ziheng Zeng, and Suma Bhat. CLCL: Non-compositional expression detection with contrastive learning and curriculum learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–743, Toronto, Canada, July 2023. Association for Computational Linguistics.