Literature Survey on Metaphor and Idiom Detection in Large Language Models

Manjusha Motamarry, Margi Shah
Natural Language Processing CS6120
December 08, 2024

Abstract:
This literature survey reviews recent advancements in metaphor and idiom detection within the context of large language models (LLMs). It examines key methods, datasets, and challenges in figurative language processing, with a focus on models like LLaMA-13B, LLaMA-30B, and GPT-3.5. The survey also explores the role of pre-trained language models (PLMs) such as BERT and RoBERTa in improving metaphor detection, as well as the integration of linguistic theories (e.g., MIP, SPV) and multi-task learning for enhanced performance. It highlights significant progress in metaphor detection but also identifies gaps in figurative language interpretation, reasoning, and generalizability.
Key findings include the importance of high-quality datasets like FLUTE, NewsMet, and MUNCH, but also the challenges these datasets face in multilingual contexts and adapting to evolving language use. While current methods excel in detecting metaphors, models still struggle with understanding deeper meanings and context, particularly in complex or novel expressions.
The survey concludes that future research should focus on developing multilingual metaphor detection models, improving contextual reasoning, and exploring unsupervised learning techniques for better model generalization. These advancements are essential for overcoming the limitations of current approaches and enhancing the practical applications of figurative language processing in areas like machine translation, sentiment analysis, and human-computer interaction.

I INTRODUCTION

1. Background:

Metaphor detection and understanding are important challenges in Natural Language Processing because metaphors enhance communication, creativity, and cultural understanding. Metaphors help explain abstract ideas by connecting them to concrete concepts. With advanced language models like BERT, RoBERTa, and GPT, the ability to interpret metaphors has improved. However, difficulties remain, especially with figurative language in different contexts like news, religious texts, and idiomatic expressions. Solving these challenges is crucial for tasks such as sentiment analysis, machine translation, and improving interactions between humans and computers.
This survey reviews research on methods and datasets designed to improve metaphor detection. It covers datasets like MelBERT[1], NewsMet[10], and MUNCH[2]. It also examines the challenges of translating idioms, and explores how metaphors are analyzed in texts like the Bhagavad Gita and the Bible. Together, these studies highlight new approaches and tools for understanding metaphors in texts.

2. Objective:

This literature survey aims to explore recent advancements in metaphor detection and understanding using machine learning and large language models. It aims to identify effective computational methods, innovative datasets, and applications in specific domains. Additionally, it seeks to address key questions, such as how state-of-the-art models handle metaphor detection across different contexts, the challenges of processing metaphors in low-resource languages, idiomatic translations, and ancient texts, and how datasets and evaluation metrics can be improved to better capture the complexity of metaphorical language.

3. Scope:

This literature survey focuses on research papers published between 2020 and 2024 to ensure relevance to recent advancements. It prioritizes high-quality contributions from peer-reviewed conferences such as ACL[11] and NAACL[12], as well as reputable journals. The survey emphasizes studies on metaphor detection, understanding, and applications in areas like news, idiomatic language, and religious or philosophical texts. It also highlights works that introduce or evaluate state-of-the-art models and datasets for metaphor processing, aiming to capture innovation in both methodologies and resources.

II METHODOLOGY

In this literature survey, a structured methodology was adopted to ensure the comprehensive identification of relevant studies in the domain of metaphor detection and figurative language understanding. The search strategy involved the use of well-established databases and search engines, including IEEE Xplore[13], Google Scholar[14], and the Association for Computational Linguistics (ACL[11]) Anthology. These platforms were selected for their extensive coverage of research in natural language processing (NLP), computational linguistics, and machine learning. A combination of key terms and phrases was employed to retrieve pertinent studies, including "metaphor detection," "figurative language understanding," "large language models (LLMs)," "natural language inference (NLI)," "metaphorical reasoning in NLP," "idiomatic expression translation," and "cross-lingual metaphor analysis." Search filters were used to refine and broaden the search as needed.

The selection of papers was based on clear inclusion and exclusion criteria. Priority was given to studies on metaphor detection, figurative language processing, and related NLP tasks. Recent research from the last five years was preferred to include the latest advancements. Only peer-reviewed journal articles, conference papers, and workshops were considered to ensure quality. Highly cited works were included to highlight influential studies, along with papers that introduced important methods or new datasets for figurative language analysis. Studies on multilingual and cross-lingual metaphor detection were also prioritized to capture the global relevance of these methods. This systematic approach provides a solid basis for understanding the current progress in figurative language processing.

II. Literature Review

1. Thematic Analysis:

1. Datasets for Figurative Language

Datasets are crucial for developing and evaluating NLP models for figurative language. FLUTE[3] offers a comprehensive dataset with 9,000 examples of figurative language, including metaphors, idioms, similes, and sarcasm, each annotated with textual explanations to support reasoning-based evaluations(Figurative). Similarly, NewsMet[10] provides a modern dataset of metaphors sourced from news headlines, emphasizing evolving and context-specific metaphorical verbs(NewsMet). The MUNCH[2] dataset focuses on metaphor interpretation through paraphrase tasks, including both suitable and unsuitable paraphrases for analysis (Metaphor Understanding). While these datasets address gaps in novelty, genre diversity, and task-specific needs, challenges remain in expanding coverage to multilingual and low-resource contexts.

2. Techniques and Models for Metaphor Detection

Recent research has introduced various computational models for metaphor detection. MelBERT[1] incorporates linguistic theories, such as the Metaphor Identification Procedure (MIP) and Selectional Preference Violation (SPV), into a contextualized BERT model, delivering strong performance on datasets like VUA and MOH-X(MelBERT Metaphor Detection). *IlliniMet*[9] enhances metaphor detection by combining RoBERTa embeddings with linguistic features like word concreteness. These methods demonstrate the effectiveness of pre-trained language models (PLMs) while also highlighting challenges, such as dependency on labeled datasets and difficulties in adapting to new contexts. A growing trend is the adoption of multi-task learning, exemplified by *GoFigure*[4], which uses idiom detection as an auxiliary task to improve metaphor detection.

3. Challenges in Figurative Language Interpretation

Interpreting figurative language remains a key challenge despite progress in metaphor detection. *Fig-QA*[7] demonstrates significant gaps in the reasoning capabilities of current language models, showing that even advanced models like GPT-3 and GPT-4 struggle with zero-shot performance on reasoning tasks involving figurative expressions. *FLUTE*[3] contributes to addressing interpretability by providing natural language explanations for figurative language, but its scope is restricted to specific categories like metaphors, idioms, similes, and sarcasm. These findings highlight the need for models that extend beyond identification to achieve robust semantic and contextual reasoning.

4. Multilingual and Domain-Specific Applications

The ability to process figurative language across languages and domains is an important area of research. Studies like *Pretrained Models for Metaphor Detection[6]* show that metaphorical knowledge in PLMs can work across different datasets and languages, especially when annotations are consistent. Research on religious texts, such as the Bhagavad Gita and the Sermon on the Mount, highlights how translation challenges affect metaphor detection in low-resource languages. These highlight the global use of NLP techniques but also show the need for better cross-lingual datasets and methods that consider cultural and domain-specific differences. Additionally, methods from *Improving LLM Abilities in Idiomatic Translation[8]* show how idiomatic knowledge bases and LLMs can preserve cultural nuances in translations, helping with challenges in low-resource settings.

Key Insights

1. Datasets: Significant progress has been made in developing high-quality datasets like FLUTE, MUNCH, and NewsMet, though gaps remain in multilingual and low-resource settings.
2. Techniques: PLMs like BERT and RoBERTa have advanced metaphor detection, but challenges in generalizability and reasoning persist.
3. Interpretation: Figurative language interpretation, especially reasoning and contextual analysis, is still an open challenge.
4. Applications: Cross-lingual and domain-specific studies reveal the potential and limitations of metaphor processing in diverse contexts.

2. Comparative Analysis

Paper Name	Methodology Used	Dataset Used	Results
FLUTE[3]	Explanation-based NLI approach with GPT-3 & crowd annotations	FLUTE	The T5 model fine-tuned on FLUTE achieved 81.8% accuracy for metaphors and 79.2% for idioms, demonstrating strong performance in figurative language detection and explanation.
GoFigure[4]	Multi-task learning combining metaphor and idiom detection	VUA, TOEFL	Achieved an F1 score of 77.5% on VUA verbs and 70.2% on TOEFL verbs, showcasing the effectiveness of multi-task learning for metaphor and idiom detection (Table 5, GoFigure!)
IlliniMet[9]	Combines RoBERTa embeddings with linguistic features like WordNet	VUA, TOEFL	Achieved an F1 score of 73.0% on all parts of speech (ALLPOS) and 77.1% on verbs in the VUA dataset. For the TOEFL dataset, it attained an F1 score of 70.3% on ALLPOS and 71.9% on verbs
MelBERT[1]	MelBERT combines linguistic theories (MIP, SPV) with BERT embeddings using a late interaction mechanism to enhance metaphor detection and generalization.	VUA, MOH-X, TroFi	MelBERT achieved an F1 score of 79.2% on the MOH-X dataset and 78.5% on VUA-18, outperforming baseline models in metaphor detection (Tables 6 and 7, MelBERT).
Interpret Figurative Language[7]	Uses the Fig-QA dataset with paired-question reasoning tasks to test language models' ability to interpret and reason about figurative language.	Fig-QA dataset	Fine-tuned RoBERTa achieved the highest accuracy of 90.32% on the Fig-QA dataset, nearing human performance (94.42%), but models still struggled with cultural metaphors.
Pretrained Models for Metaphor Detection[6]	Tests pre-trained language models (BERT, RoBERTa, ELECTRA) for metaphor detection using various datasets and cross-lingual tasks.	VUA, TroFi, LCC	ELECTRA achieved the highest accuracy of 89.3% on LCC and 83.03% on VUA POS, demonstrating superior metaphor detection and cross-dataset transferability compared to other PLMs.

Improving LLM Abilities in Idiomatic Translation[8]	Cosine similarity lookup and idiom matching techniques	Custom dataset (Urdu idioms, IdiomKB)	Cosine Similarity Lookup outperformed direct translations with GPT-4o achieving an average score of 2.761 for ZH → EN and 2.629 for EN → UR, demonstrating effectiveness in preserving idiomatic meanings.
Bhagavad Gita and Sermon on the Mount[5]	Analyzes translation and metaphor detection in low-resource texts	Bhagavad Gita, Sermon on the Mount	The pre-trained LLM framework effectively identified complex metaphors in translations of religious texts, with qualitative analysis confirming fair consistency across translations despite differing vocabularies
MUNCH[2]	Paraphrase generation and evaluation for metaphor interpretation	MUNCH	Models like GPT-3.5 achieved a top recall of 32% and a mean reciprocal rank of 0.54 on the MUNCH dataset, highlighting challenges in metaphor understanding and paraphrase reasoning.
NewsMet[10]	The paper uses a custom-annotated dataset of news headlines to fine-tune RoBERTa for metaphor detection in domain-specific contexts.	NewsMet	RoBERTa fine-tuned for metaphor detection achieved F1 scores of 78% on literal and 76% on metaphorical sentences.

The studies highlight various methods and progress in figurative language processing. *FLUTE[3]* showed strong results with 81.8% accuracy for metaphors and 79.2% for idioms, while *GoFigure[4]* used multi-task learning to achieve F1 scores of 77.5% on VUA verbs and 70.2% on TOEFL verbs. *IlliniMet[9]* and *MelBERT[1]* combined linguistic features with PLMs, achieving F1 scores of 77.1% and 79.2% on VUA verbs and MOH-X, respectively. The Fig-QA dataset showed fine-tuned RoBERTa achieving 90.32% accuracy, close to human performance, though cultural metaphors remain a challenge. ELECTRA performed best on multiple datasets, with 89.3% accuracy on LCC. *Improving LLM Abilities in Idiomatic Translation[8]* succeeded in preserving idiomatic meanings using Cosine Similarity Lookup. The *Bhagavad Gita and Sermon on the Mount[5]* study found fair consistency in metaphor detection across translations. *MUNCH[2]* highlighted challenges in paraphrase reasoning, with GPT-3.5 achieving a recall of 32%. Finally, *NewsMet[10]* showed RoBERTa's effectiveness in detecting metaphors in news headlines, with F1 scores of 78% for literal and 76% for metaphorical sentences. These studies show progress but also highlight challenges like reasoning, cultural understanding, and adaptability across domains.

The conclusions of the papers highlight key progress and challenges in figurative language processing across datasets. Studies like *FLUTE[3]*, *GoFigure[4]*, and *MUNCH[2]* stress the importance of reasoning-based evaluations and task-specific datasets to improve metaphor and idiom understanding. *NewsMet[10]* and *Pretrained Models for Metaphor Detection[6]* show that domain-specific and multilingual datasets are helpful but face issues with inconsistent annotations. *IlliniMet[9]* and *MelBERT[1]* demonstrate that combining linguistic theories with PLMs improves metaphor detection, though reasoning and generalization remain challenges. The *Bhagavad Gita* and *Sermon on the Mount[5]* study highlights difficulties in detecting metaphors across translations due to stylistic and vocabulary differences. Overall, while datasets have improved metaphor detection, reasoning, scalability, and cross-lingual adaptation remain key challenges for future research.

Strengths and Weaknesses of Approaches

Paper Name	Strengths	Weaknesses
FLUTE[3]	Improves reasoning with explanations; covers various figurative types.	Focuses on only four types of figurative language.
GoFigure[4]	Uses extra tasks like idiom detection to improve results.	Complex and harder to scale for larger tasks.
IlliniMet[9]	Combines advanced models with extra language tools like WordNet.	Heavily depends on external tools and resources.
MelBERT[1]	Blends linguistic theories with modern models for accuracy.	Good at detecting metaphors but not at deeper understanding.
Fig-QA[7]	Points out reasoning gaps in language models.	Models still struggle with reasoning and zero-shot tasks.
Pretrained Models for Metaphor Detection[6]	Shows metaphor knowledge can transfer between languages.	Inconsistent annotations can cause issues.
Improving LLM Abilities in Idiomatic Translation[8]	Keeps cultural and idiomatic meanings intact in translations.	Small dataset limits results, especially for rare languages.
Bhagavad Gita and Sermon on the Mount[5]	Studies metaphor translation in less common languages.	Differences in translations can reduce accuracy.
MUNCH[2]	Focuses on paraphrasing metaphors for better understanding.	Mainly works on paraphrasing tasks, missing broader insights.
NewsMet[10]	Provides a modern dataset for analyzing metaphors in news.	Limited to news headlines; lacks variety in text types.

The strengths and weaknesses of these approaches reveal important trends in figurative language research. A key strength is the use of linguistic theories and advanced models, as seen in *MelBERT[1]* and *IlliniMet[9]*, which achieve high accuracy in metaphor detection. Multi-task learning in *GoFigure[4]* and reasoning-focused tasks in *MUNCH[2]* and *Fig-QA[7]* improve understanding and interpretation. However, many methods rely heavily on labeled datasets, making it harder to adapt to new contexts and languages. Cross-lingual approaches like *Pretrained Models for Metaphor Detection[6]* and *Idiomatic Translation[8]* show promise but face challenges in low-resource settings. While there has been significant progress, issues with scalability, reasoning, and multilingual support remain.

Conclusion for Comparative Analysis

In conclusion, the comparative analysis highlights significant progress in figurative language processing through diverse methods, datasets, and models. Detection-focused approaches like *MelBERT[1]* and *IlliniMet[9]* excel in accuracy, while reasoning-based models such as *MUNCH[2]* and *Fig-QA[7]* address interpretation challenges but reveal gaps in nonliteral reasoning. Domain-specific datasets like NewsMet[10] and *FLUTE[3]* enhance metaphor detection, and cross-lingual studies like Pretrained Models and Idiomatic Translation demonstrate potential for multilingual applications. However, challenges in reasoning, scalability, and adaptability across languages and contexts persist, pointing to the need for further research to build more robust and versatile models.

III Critical Analysis

1. Evaluation of Research Quality

The research in figurative language processing is of high quality, with many studies advancing metaphor detection and understanding. Most methods use pre-trained language models (PLMs) like BERT and RoBERTa, which are good at capturing contextual meaning—a key strength in language tasks. Innovative approaches, such as integrating linguistic theories (e.g., MIP and SPV) in *MelBERT[1]* and multi-task learning in *GoFigure[4]*, have achieved strong results on datasets like VUA and MOH-X. Models like *FLUTE[3]*, which include natural language explanations, improve the evaluation of both detection and reasoning. Datasets like *NewsMet[10]* and *MUNCH[2]* provide real-world examples of metaphors, adding value to the field. However, these models often rely on labeled datasets and struggle to adapt to new or unseen contexts. Even advanced PLMs like GPT-3 and GPT-4, while strong in many tasks, still face difficulties in reasoning and understanding figurative language, showing that they fall short of human-level comprehension.

This research has both practical and theoretical importance. Better metaphor detection improves natural language understanding (NLU) and is essential for applications like machine translation, sentiment analysis, and chatbots, where figurative language is common. Models like *FLUTE[3]*, which focus on interpretability, make NLP systems more transparent and reliable, improving human-AI collaboration in areas like healthcare, education, and customer service. The research also helps us understand how humans process metaphors and figurative language, bridging the gap between linguistic theories and computational methods. These findings underline the need for continued innovation in this area.

2. Identification of Gaps

Despite progress, gaps remain in metaphor detection in Large language models:

1. Cross-lingual and Multilingual Models: While some studies explore metaphor detection in different languages, multilingual metaphor processing is underdeveloped. More diverse datasets are needed for non-English languages and low-resource settings.
2. Reasoning and Contextual Understanding: Models still struggle with understanding the deeper meaning and context of figurative language. Future research should focus on improving reasoning abilities in metaphor interpretation.
3. Dynamic Language Use: Many datasets, like VUA, focus on older language and do not reflect how metaphors evolve in modern language, particularly in digital communication and social media.
4. Generalizability: Many models rely on labeled data and struggle to adapt to new contexts. Future work should explore unsupervised or zero-shot learning techniques for more adaptable models.

3. Implications

1. Practical Implications:
   The findings have several real-world applications. In machine translation, improved metaphor detection can lead to more accurate translations, especially for idiomatic expressions. In sentiment analysis and social media monitoring, a better understanding of figurative language can improve the detection of sarcasm, irony, and humor. Virtual assistants and chatbots can also benefit, making interactions more natural and human-like.
2. Theoretical Implications:
   The theoretical implications of this research include improving our understanding of figurative language and its connection to linguistic theories. By using frameworks like MIP and SPV in machine learning models, the research helps simulate how humans understand metaphors. While it doesn't directly study how the brain works, it offers valuable insights into metaphor processing and supports the development of NLP systems that handle complex, context-dependent language. It also adds to conceptual metaphor theory and linguistics by providing computational perspectives on figurative language.

4. Limitations

While the studies reviewed offer valuable insights, there are several limitations:

1. Dataset Bias: Datasets like VUA and NewsMet[10] focus on specific types of figurative language (e.g., metaphors, idioms) and are often limited to English, reducing their ability to capture diverse expressions across languages andE contexts.
2. Domain-Specific Focus: Some datasets, such as NewsMet[10], are limited to news headlines, making them less applicable to other text types like literature or conversational speech.
3. Interpretation Challenges: Despite progress in metaphor detection, models still struggle with interpreting complex or novel figurative language, making deeper semantic understanding a key challenge.
4. Limitations of the Review: This review may not cover all relevant studies, particularly those published after the review period. Additionally, it focuses on PLMs and linguistic theories, potentially overlooking other promising approaches, like neural-symbolic models or hybrid systems.

5. Conclusion for the Critical Analysis

In conclusion, the reviewed studies show significant progress in metaphor detection and figurative language understanding using pre-trained language models. However, challenges remain in reasoning, generalization, and multilingual applications. Future research should focus on improving contextual reasoning, developing cross-lingual datasets, and exploring unsupervised learning methods. The practical applications are wide-ranging, from machine translation to virtual assistants, while the theoretical insights contribute to cognitive science and linguistics. Despite its limitations, this research provides a solid foundation for further advancements in figurative language processing.

IV. Conclusion

1. Summary of Findings

The literature reviewed shows significant progress in figurative language processing, especially in detecting and interpreting metaphors. Pre-trained language models (PLMs) like BERT and RoBERTa are effective tools for metaphor detection, improving accuracy and contextual understanding. Several innovative approaches, such as integrating linguistic theories (e.g., MIP and SPV), using multi-task learning, and adding natural language explanations for interpretability, have been explored.

While current methods excel in metaphor detection, there are still gaps in figurative language interpretation, reasoning, and generalizability. Datasets like FLUTE[3], NewsMet[10], and MUNCH[2] provide valuable data, but challenges remain in multilingual contexts and adapting to evolving languages. Models continue to struggle with reasoning about the meaning and context of figurative expressions, especially in complex or novel cases.

2. Future Directions

Future research should focus on improving multilingual metaphor detection, and developing models that can handle figurative language in non-English languages, particularly in low-resource settings. Additionally, reasoning and contextual understanding remain significant challenges, so future models should incorporate semantic reasoning capabilities to understand the deeper meanings of figurative expressions. Finally, enhancing model generalization through transfer learning or few-shot learning can improve the adaptability of models to new contexts with minimal labeled data. Addressing these gaps will be key to advancing figurative language processing.

V References

[1] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., & Lee, J. (2021). MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1763–1773. https://doi.org/10.18653/v1/2021.naacl-main.141

[2] Tong, X., Choenni, R., Lewis, M., & Shutova, E. (2024). Metaphor understanding challenge dataset for LLMS. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3517–3536. https://doi.org/10.18653/v1/2024.acl-long.193

[3] Chakrabarty, T., Saakyan, A., Ghosh, D., & Muresan, S. (2022). FLUTE: Figurative Language Understanding through Textual Explanations. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2022.emnlp-main.481

[4] Chen, X., Leong, C. W., Flor, M., & Beigman Klebanov, B. (2020). Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS Team in 2020 Metaphor shared task. *Proceedings of the Second Workshop on Figurative Language Processing*, 235–243. https://doi.org/10.18653/v1/2020.figlang-1.32

[5] Chandra, R., Tiwari, A., Jain, N., & Badhe, S. (2024). Large language models for metaphor detection: Bhagavad Gita and Sermon on the Mount. *IEEE Access*, *12*, 84452–84469. https://doi.org/10.1109/access.2024.3411060

[6] Aghazadeh, E., Fayyaz, M., & Yaghoobzadeh, Y. (2022). Metaphors in Pre-Trained Language Models: probing and generalization across datasets and languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.144

[7] Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022). Testing the ability of language models to interpret figurative language. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4437–4452. https://doi.org/10.18653/v1/2022.naacl-main.330

[8] Donthi, S., Patel, O., Rohan, E., Spencer, M., Doh, J., Zhu, K., & O'Brien, S. (2024). *Improving LLM Abilities in Idiomatic Translation*. https://doi.org/ https://doi.org/10.48550/arXiv.2407.03518

[9] Gong, H., Gupta, K., Jain, A., & Bhat, S. (2020). Illinimet: Illinois system for metaphor detection with contextual and linguistic information. *Proceedings of the Second Workshop on Figurative Language Processing*, 146–153. https://doi.org/10.18653/v1/2020.figlang-1.21

[10] Joseph, R., Liu, T., Ng, A. B., See, S., & Rai, S. (n.d.). NewsMet : A 'do it all' Dataset of Contemporary Metaphors in News Headlines. *Findings of the Association for Computational Linguistics: ACL 2022*, 10090–10104. https://doi.org/10.18653/v1/2023.findings-acl.641

[11] https://aclanthology.org/venues/acl/

[12] https://aclanthology.org/venues/naacl/

[13] https://ieeexplore.ieee.org/Xplore/home.jsp

[14] https://scholar.google.com/