Cracking the Figurative Code: A Survey of Metaphor Detection Techniques

Vrinda Kohli[1], Himanshu Nandanwar[2] and Rahul Katarya[3]

[1] Manipal University Jaipur, India
vrinda.209301473@muj.manipal.edu
[2] Delhi Technological University, India
himanshunandanwar9cm0@gmail.com
[2] Delhi Technological University, India
rahuldtu@gmail.com

Abstract

Metaphor Detection is a crucial area of study in computational linguistics and natural language processing, as it enables the understanding and communication of abstract ideas through the use of concrete imagery. This survey paper aims to provide an overview of the current state-of-the-art approaches that tackle this issue, and analyze trends in the domain across years.
The survey recapitulates the existing methodologies for metaphor detection, highlighting their key contributions and limitations. The methods are assigned three broad categories, namely feature-engineering based, traditional deep learning-based, and transformer-based approaches. An analysis of strengths and weaknesses of each category is showcased.
Furthermore, the paper explores the annotated corpora that have been developed to facilitate the development and evaluation of metaphor detection models. By providing a comprehensive overview of the work already done and the research gaps present in pre-existing literature, this survey paper aims to help future research endeavors, and thus contribute to the advancement of metaphor detection methodologies.

Keywords: Metaphor Detection, Natural Language Processing, Linguistic Analysis, Computational Linguistics, Lexical Semantics

1 Introduction

Roughly 12% of the words used in a natural language document are used metaphorically [1]. Metaphors are linguistic tools that present comparisons between two seemingly unrelated ideas through shared traits. They act as a means to describe abstract concepts through vivid imagery. A metaphor is defined by a stark difference in its literal and contextual meanings (Fig 1). For example, in the phrase “I am a forest fire” [2], the speaker does not actually mean that she is a forest fire, but instead uses the phrase to convey the raging intensity of her emotions, displaying a vast disparity between the literal and contextual sense of the expression “forest fire”.

Automated Metaphor Detection boils down to identification of a metaphorical word (or token) in a given text sequence by a machine learning model. This demands a deeper understanding of the often subtle, figurative language used which requires computational models to go beyond surface-level interpretations and delve into the underlying semantic layers of the sentence in order to capture relevant contextual information. Consequently, the detection of metaphors warrants sophisticated approaches that can encompass the intricacies in the interplay between language, context, and figurative expressions to achieve reliable and insightful results. This task also shows importance in other natural language processing tasks such as machine translation [3], sentiment analysis or opinion mining [4], dialogue systems [5] and machine reading comprehension [6].

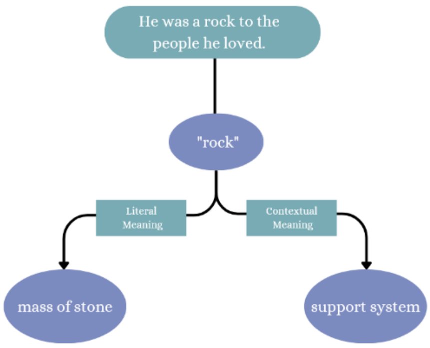

Fig. 1. Metaphors have different literal and contextual meanings.

The pre-existing techniques for metaphor detection can be broadly classified into three categories. Feature based methodologies deal with extracting metaphor specific features from the corpus to identify the needed. Traditional Deep Learning based approaches employ various RNN and hybrid architectures to model the sequential nature of sentences. Lastly, transformer-based approaches use attention equipped encoder-decoder style pretrained architectures (BERT, RoBERTa etc.) to capture semantic and syntactic relationships from the input text.

Thenceforth, the study of metaphor detection holds considerable implications for understanding language, cognition, and communication. By examining the existing literature, this survey paper attempts to shed a light on research gaps. This paves a way for further advancements in the field for developing robust and context-aware models that show generalization across different languages, cultures, and domains. Through this paper, we hope to provide a comprehensive resource for researchers interested in the field of automated metaphor detection.

2 Literature Review

The techniques employed for metaphor detection (MD) have witnessed various trends over the years. In the earlier years of research about this problem, a lot of focus was given to hand-crafted metaphor-centric features. [7] used word concreteness and abstractness as a defining feature, while [8] used feature norms. Imaginability [9], bag-of-words features [10] and sparse distributional features [11] have also been used as linguistic features for machine learning models.

Next came techniques utilizing Neural architectures, such as BiLSTM [12], CNN-hybrids [13] and Graph Neural Networks [14] [15]. These methods popularized the use of word embeddings such as GloVe [16] and Elmo [17] vectors for metaphor detection. [18] further integrates linguistic theory conventions Metaphor Identification Procedure (MIP) [19] and Selectional Preference Violation (SPV) [20] by modeling them as neural architectures.

Transformer based approaches typically model linguistic rules and other contextual information by using BERT or RoBERTa encoder modules, using those in conjunction with techniques such as context denoising [21], self-supervised learning [22], reading comprehension [23] and parse-tree alterations [14].

A detailed survey covering the specifications of all three approaches can be found in Table-1, and Table-2 demonstrates the quantifiable results obtained by these models.

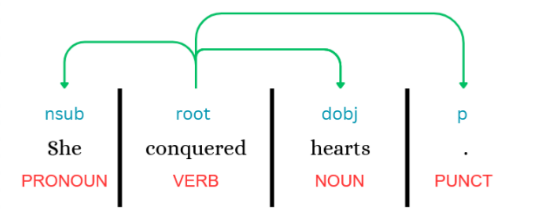

Fig. 2. Metaphors with verb-noun direct object relation

2.1 Publicly Available Datasets

There are primarily three datasets on which experimentation pertaining to MD tasks is performed.

VUA: The VU Amsterdam Metaphor Corpus (VUA) [24] dataset is the largest publicly available dataset annotated for metaphor detection tasks. It is sampled from the

British National Corpus across four genres (Academic, News, Conversation, and Fiction), and consists of 117 fragments. It has over 2K unique verbs, and the metaphors are distributed with natural likelihood (~10%).

MOH-X: MOH-X [25] is a verb metaphor detection dataset that has datapoints sampled from WordNet [26] example sentences. Each sentence has only a singular metaphor tagged in it. The average sentence length is 8 tokens and 48.69% of the words are metaphorical in nature.

TroFi: TroFi [27] is a single target verb metaphor detection dataset which is comprised of sentences from 1987-1989 Wall Street Journal Corpus Release-1. The average length for this dataset is 28.3 tokens per sentence, which is the longest among the three datasets explored. The percentage distribution of metaphors in the dataset amounts to 43.54%.

Table 1. Existing Methodologies

Model	Year & Ref	Category	Contribution	Methodology	Limitations	Advantages
BiLSTM	2018 [12]	Traditional DL approach	Utilization of BiLSTM models with ELMo embeddings for MD.	Tokens concatenated with their ELMo embeddings are encoded using a BiLSTM module. The detection task is modelled in two ways: the classification task is done by using a feedforward neural network, and the sequence labelling task applies an attention layer for computing attention weight per token for weighted classification.	BiLSTM encoder struggles in capturing metaphors with long-range dependencies, indirect metaphors and personification related metaphors.	Infers that predicting metaphor labels of context words helps predict the target word and that contextualized word vectors improves model performance
Disc	2019 [28]	Feature Engineering approach	Usage of broader discourse-based features to train gradient boosting classifiers for MD task	The GloVe embeddings, doc2vec vectors, skip-thought vectors and ELMo embeddings are obtained and their concatenation is used as a feature-vector for an input to a gradient boosting algorithm (XGBoost)	Conversation based metaphors are harder to detect and this approach has an a-priori need for broader-context beyond sentence level.	Competitive results without neural architectures or manually-engineered metaphor specific features. The usage of paragraph level context vastly improves detection performance.

DeepMet	2020 [23]	Transformer based approach	Reading comprehension paradigm for MD at a token level.	MD is considered to be a reading comprehension task, based on context and query words. It involves inputting global and local text contexts, query features, POS features and FGPOS features into a Siamese architecture with two separate BERT encoders for local and global features. The encoders share weights and an average pooled vector is used as input to the metaphor discrimination module. Cross validation introduces a metaphor preference parameter.	Faces difficulties in detecting metaphors triggered by multiple words since the queries are answered one word at a time. Downsampling via average pooling may lead to loss of relevant information.	Demonstrates that FGPOS features provide more information than standard POS features. The metaphor preference parameter models real world scenario in its dealing with imbalanced datasets.
WSD-GCN	2020 [14]	Traditional DL approach	Leverages Graph Convolution Networks (GCN) with dependency parse trees and a multi-task framework for exploiting the similarity of MD and word sense disambiguation (WSD) task.	A BiLSTM is used to obtain a feature vector from GLoVe, ELMo and index embeddings of the sentence, which is then inputted into a GCN module. The GCN and BiLSTM vectors are aggregated via calculated control vectors that filter out irrelevant information. A dense network with a Softmax layer is used for MD. Owing to the multi-task approach. Two encoders are trained alternatively and simultaneously for WSD and MD to share knowledge between the two tasks.	The usage of dependency parse trees imposes a reliance on the dataset structure for successful generalization of the approach. A lack of cross dataset evaluation leaves the question of generalizability unanswered. This technique is hard to apply to batch-optimization due to complicated tree-related structure.	The GCN approach successfully identifies relevant context words based on their importance. The multi-task approach handles the issue of knowledge transfer between two tasks when the dataset is only annotated for one of the two.
MWE-GCN	2020 [15]	Traditional DL-based approach	Introduces a multiword expression aware model for metaphor identification	The Dependency parse tree information is treated as an undirected graph. The adjacency matrix of this graph is linearly	No comparison with the standard VUA dataset, which is considerably vast in its information and generalization	Demonstrates that the knowledge of Multiword Expressions can significantly boost the

				combined with attention-based matrices, providing fully connected weighted graph matrices to determine relation strength between nodes. These matrices are inputted to different Graph Convolution Networks, the outputs from which are linearly combined. The same process is followed for token-level relations between multiword expression components present in the sentence. The GCN outputs of both architectures are concatenated and passed through another GCN to obtain results.	strength is not evaluated. The complex tree-related structure makes this approach less amenable to batch optimization.	performance of MD methods
MelBERT	2021 [1]	Transformer based approach	Uses contextualized word representations and linguistic theories, namely Metaphor Identification Protocol (MIP) and Selectional Preference Violation (SPV) for MD	SPV and MIP are modelled using two RoBERTa backboned encoders and a combined prediction score is obtained post late-stage interaction.	Borderline or implicit metaphors are much harder to identify. The syntactic structure isn't utilized as context words across sub-sentences lose their relation.	Since late interactions are utilized between the two lingual rules, the sentence vectors can be reused, leading to an amortized cost of encoding. A good level of generalization is achieved across datasets as exhibited in Zero Shot experimentation.
CATE	2021 [22]	Transformer based approach	Introduces a semi-supervised self-training strategy for collecting large-scale candidate instances from generated unlabeled corpus, and a contrastive objective for	A BERT model is finetuned using pre-existing labelled data. A Target-based Generating Strategy is used to create a large-scale, relevant unlabeled corpus. The finetuned model pseudo-labels this corpus, and this data is then used to augment the training data.	When the available training data size is high, the net gain from self-training drops. Model accuracy drops when words from multiword expressions are utilized in their literal sense.	Significant improvement when small-scale datasets are used due to self-supervised data augmentation. Self-training leads to a more diverse dataset, bringing about better MD in underrepresented genres. The contrastive objective quantifies

			capturing MIP is defined.	The fine-tuned model is updated iteratively using a self-training strategy.		the contrast between literal and contextual meanings, upholding MIP without a bulky architecture.
CIA*	2022 [29]	Feature Engineering Approach	Lightweight algorithm for Direct Object related metaphors (Fig 2) specific to the cybersecurity domain	Bing API is queried for top 50 websites related to a selected verb, relevant sentences are extracted and added to the corpus which is then parsed to obtain collocated nouns. The synsets and hyponyms for these nouns are obtained via WordNet. If the main synset is not present in the collocated nouns list, the word is predicted to be a metaphor.	Only a particular style of metaphors is evaluated, constricting the extent of evaluation.	Comparable results without bulky deep learning architecture. The development of a real-world corpus is simple enough to be extended for usage across multiple domain-specific tasks. This approach can identify multiple metaphorical instances present in a sentence successfully.
Frame-BERT	2023 [31]	Transformer based approach	Explainable and interpretable metaphor detection by incorporating FrameNet embeddings.	Two RoBERTa encoders are used: the conceptual encoder processes the FrameNet embeddings and the sentence encoder models MIP and SPV. The outputs from both encoders are concatenated to obtain input for classification module.	Features such as Frame Elements, Lexical Units and context graphs need to be explored.	Usage of FrameNet embeddings brings up performance by 1.2% owing to their ability to capture deep-level semantics.
RoPPT	2023 [21]	Transformer based approach	A target-oriented parse tree structure is utilized for MD by extracting semantically relevant neighbors of a target word.	The original parse tree is reshaped by rooting the tree at the target word. Context Denoising is performed by pruning the tree based on the distance between the root and leaves. Two RoBERTa based encoders are used for encoding, one for the target word, and the other for the input sentence, followed by a classification module.	The usage of average pooling may lead to loss of fine-grained details. Performance is lower than expected for shorter sentences.	The modified tree structure allows the model to focus on only relevant information with regard to the target word. Irrelevant parts are ignored despite their position in the input sentence. Demonstrates the robustness of context denoising mechanism over long sentences.

Table 2: Results on various metrics

Ref	Model	VUA				TroFi				MOH-X			
		P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
[1]	MelBERT	80.1	76.9	78.5	-	53.4	74.1	62.0	-	79.3	79.7	79.2	-
[12]	BiLSTM	68.2	71.3	69.7	81.4	70.7	71.6	71.1	74.6	79.4	73.5	75.6	77.2
[14]	WSD-GCN	74.8	75.5	75.1	93.8	73.1	73.6	73.2	76.4	79.7	80.5	79.6	79.9
[15]	MWE-GCN	-	-	-	-	73.78	71.81	72.78	73.45	79.98	80.40	80.19	80.47
[21]	RoPPT	80.0	78.2	79.1	-	54.2	76.2	63.3	-	77.0	83.5	80.1	-
[22]	CATE	79.3	78.8	79.0	94.8	74.4	74.8	74.5	77.7	85.7	84.6	84.7	85.2
[23]	DeepMet	75.6	78.3	76.9	91.6	72.1	80.6	76.1	77.0	93.3	90.3	91.8	92.3
[28]	Disc	58.9	77.1	66.8	-	-	-	-	-	-	-	-	-
[29]	CIA*	-	-	-	-	72	66	68	69	-	-	-	-
[30]	Frame-BERT	82.7	75.3	78.8	-	70.7	78.2	74.2	-	83.2	84.2	83.8	-

3 Research Gaps

After a thorough analysis of existing works, as shown in Table-1, we have identified the challenges and limitations of prior approaches as follows:

3.1 Low Generalizability

On an average, the proposed approaches rarely discuss the generalizability across datasets, barring a few exceptions [1] [31]. Probing based studies done in [32] demonstrate that there are large gaps present between the in-distribution and out-of-distribution performances of Transformer based methods for MD tasks, presumably due to annotation bias present across the datasets. This implies that the generalizability across datasets of such approaches is lower than expected.

3.2 Heavy Dependency on Dataset

Upon analyzing trends across various methods, one common denoting factor is that these techniques are highly dataset specific, which poses as a challenge for generalization on real-world data which is usually much more diverse in its linguistic styles, cultural references and domain-specific terminologies. There is a need to develop methods which do not depend this heavily on their training corpus.

3.3 LLM-centric Approaches

[14] shows competitive results in MD tasks by leveraging its similarity to Word Sense Disambiguation (WSD) [33]. It is shown in [34] the successful usage of LLMs for solving the WSD task. Thus, cross-domain knowledge can be utilized to apply similar

techniques for LLM centric approaches for MD.

4 Discussion

There are primarily three categories of methodologies discussed in this survey, each having its own inherent drawbacks and benefits. Even though all methods show a certain level of sensitivity towards the corpus quality, these effects are vastly pronounced in Feature Engineering based methods. These methods are only as good as the hand-crafted features utilized by them and the process of extracting corpus-specific features implies a lack of generalization capability across unseen data. Thus, rarely used metaphors are difficult to identify [1].

Traditional deep learning-based approaches often lack interpretability. Due to the shallow nature of the neural architectures used, the entire extent of context information across different hierarchical levels is not obtained [23].

Transformer based methodologies were proposed to primarily tackle the limitations induced by shallowness of these methods. Due to their superior ability to encode metaphorical knowledge [32] these show state-of-the-art performance on MD tasks (Table-2).

5 Conclusion

Summing up, a number of approaches broaching automated detection of metaphors in natural language corpora were discussed in this paper. We have discussed the linguistic aspects of metaphor and how they get modeled as computational tasks. Understanding and recognizing metaphors rigorously through computational techniques is bound to bring significant progress in not only the aligned natural language processing tasks but also provide an insight into human cognition.

As the field continues to advance, researchers should focus on developing robust and context-aware models that tackle the prevalent issues with prior techniques, integrating up-and-coming innovations within them. A possible course of action for the authors would be to explore and apply themselves to the research gaps and look into LLM-based methodologies for metaphor detection.

In conclusion, by providing a thorough understanding of the current landscape, challenges, and limitations of the current methods for metaphor detection, this paper hopes to facilitate future research endeavors and foster collaborative efforts for development of advanced metaphor detection techniques.

References

1. Minjin Choi: MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics. (2021)

2. Mitski: A Burning Hill [Song] on *Puberty 2.* Dead Oceans.
3. Chunqi Shi: Translation agent: A new metaphor for machine translation. In: Proceedings of New Gener. Comput., 32(2):163–186 (2014)
4. Erik Cambria: Sentiment analysis is a big suitcase. In: Proceedings of IEEE Intell. Syst., 32(6):74–80 (2017)
5. Pawel Dybala: Humor, emotions and communication: Human-like issues of human-computer interactions. In: Proceedings of Annual Meeting of the Cognitive Science Society (CogSci) (2012)
6. Ming Tu: Multi-hop Reading Comprehension with Multiple Heterogeneous Tasks. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics. (2019)
7. Peter D. Turney: Literal and metaphorical sense identification through concrete and abstract context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 680–690 (2011)
8. Luana Bulat: Modelling metaphor with attribute-based semantics. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 523–528 (2017)
9. George Aaron Broadwell: Using Imageability and Topic Chaining to Locate Metaphors in Linguistic Corpora. In: Proceeding of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP), pages 102–110. (2013)
10. Maximilian Köper : Distinguishing literal and non-literal usage of German particle verbs. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics. (2016)
11. Chee Wee Leong: A report on the 2018 VUA metaphor detection shared task. In: The Workshop on Figurative Language Processing (FigLang@NAACL-HLT), pages 56–66 (2018)
12. Ge Gao: Neural Metaphor Detection in Context. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics (2018)
13. Chuhan Wu: Neural Metaphor Detecting with CNN-LSTM Model, In: *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics. (2018)
14. Duong Minh Le: Multi-Task Learning for Metaphor Detection with Graph Convolutional Neural Networks and Word Sense Disambiguation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8139-8146. (2020)
15. Omid Rohanian: Verbal Multiword Expressions for Identification of Metaphor. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2890–2895, Online. Association for Computational Linguistics (2020)
16. Jeffrey Pennington: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing, pages 1532–1543
17. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1, pages 2227–2237 (2018)
18. Rui Mao: End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories, In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics. (2019)
19. Pragglejaz Group: MIP: A method for identifying metaphorically used words in discourse. Metaphor and symbol, 22(1):1–39. (2007)

20. Yorick Wilks: A preferential, pattern-seeking, semantics for natural language inference. Artificial intelligence, 6(1):53-74 (1975)
21. Shun Wang: Metaphor Detection with Effective Context Denoising. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1404–1409, Dubrovnik, Croatia. Association for Computational Linguistics. (2023)
22. Zhenxi Lin: CATE: A Contrastive Pre-trained Model for Metaphor Detection with Semi-supervised Learning. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3888–3898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. (2021)
23. Chuandong Su: DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics. (2020)
24. Gerard J Steen: A method for linguistic metaphor identification: From MIP to MIPVU, volume 14. John Benjamins Publishing. (2010)
25. Saif Mohammad: Metaphor as a medium for emotion: An empirical study. In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, pages 23–33 (2016)
26. Christiane Fellbaum: WordNet: An Electronic Lexical Database. Bradford Books. (1998)
27. Julia Birke: A clustering approach for nearly unsupervised recognition of non-literal language. In 11th Conference of the European Chapter of the Association for Computational Linguistics. (2006)
28. Jesse Mu: Learning Outside the Box: Discourse-level Features Improve Metaphor Identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 596–601, Minneapolis, Minnesota. Association for Computational Linguistics. (2019)
29. Kelsey Hinton: Metaphor identification in cybersecurity texts: a lightweight linguistic approach. *SN Appl. Sci.* **4**, 60. https://doi.org/10.1007/s42452-022-04939-8 (2022)
30. Yucheng Li: FrameBERT: Conceptual Metaphor Detection with Frame Embedding Learning. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia. Association for Computational Linguistics (2023)
31. Shenglong Zhang: Metaphor Detection via Linguistics Enhanced Siamese Network. In: *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. (2022)
32. Ehsan Aghazadeh: Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 2037 – 2050 (2022)
33. Eneko Agirre: Word Sense Disambiguation using Conceptual Density. In: Proceedings of 1st International Conference on Recent Advances in Natural Language Processing, Velingrad (1995)
34. Yurii Laba: Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation. In: *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 11–19, Dubrovnik, Croatia. Association for Computational Linguistics. (2023)