Metaphor Detection and Translation for Representing Natural Language in Formal Logic

Jarrad Singley[1 [0009−0009−7640−3782]], Adam Pease[1 [0000−0001−9772−1266]], and Richard Thompson[1 [0009−0001−6541−1092]]

Naval Postgraduate School, Monterey, CA, USA
{jarrad.singley, richard.thompson, adam.pease}@nps.edu

Abstract. Transforming natural language into formal logic statements enables machines to reason over a curated ontology. Metaphors pose a particular problem for automated translation and prevent reliable translation. This research presents a novel resolution framework that preprocesses natural language, removing and replacing non-literal words and phrases with literal, metaphor-free interpretations.
Our method adapts a transformer based metaphor detector trained on a benchmark metaphor dataset, and then generates batches of candidate translations using a large language model (LLM). The best candidate is selected based on its similarity to the original sentence and the presence of residual metaphors detected. Performance was evaluated by re-detecting metaphors in the output and using a standard dataset of metaphorical-literal sentence pairs to measure similarity between the module's output and the dataset's literal sentence. Our method demonstrated marked improvement in the state of the art, significantly increasing the success rate at which metaphors were reduced in natural language.

Keywords: Metaphor detection techniques · metaphor translation · metaphor interpretation · language-to-logic translation · automated language formalization · SUO-KIF logic · conceptual metaphors · large language models (LLMs) · Hybrid Neuro-Ontological Language Understanding (HyNOLU) · sentence similarity

1 Introduction

Modern artificial intelligence (AI) systems, though powerful in natural language processing, face challenges of trustworthiness [1], logical reasoning [2], and interpretability [3]. While they generate fluent text, they are prone to errors, hallucinations, and contradictions, making them unreliable in domains requiring correctness and justification.

The Hybrid Neuro-Ontological Language Understanding (HyNOLU) project [4] addresses these challenges by developing a pipeline for converting natural language into the formal language Standard Upper Ontology–Knowledge Interchange Format (SUO-KIF) [5] using logical symbols from the Suggested Upper Merged Ontology (SUMO)[1] [6, 7]. Formal language enables reasoning with automated theorem provers such as Eprover [8] and Vampire [9]. By combining language model-based translation with formal logic and automated theorem proving, its goal is to produce representations that can be logically tested, allowing machines to reason, detect contradictions, and deliver verifiable conclusions to human analysts. The HyNOLU architecture employs a modular pipeline that integrates pre-processing, language to logic (L2L) translation, and automated theorem proving (ATP)-based validation to convert natural language into SUO-KIF. Within the pre-processing component, metaphor detection and resolution plays a crucial role in removing figurative language, ensuring that downstream logic translation operates on precise and literal meaning.

This paper focuses on this metaphor resolution stage of the HyNOLU architecture, which is critical for enabling downstream modules to operate on literal, logically tractable inputs. To support this task, a transformer-based metaphor detector [10] is adapted and integrated into the translation process. The detector provides token-level metaphor flags that enable finer-grained control over which words require literal substitution, allowing sentences without metaphors to bypass processing while those containing figurative content receive targeted resolution.

This work contributes a batch generation method that produces multiple candidate translations and selects among them using a joint criterion of token-level metaphor re-detection and sentence similarity. This selection strategy ensures that the chosen output minimizes metaphorical content while maintaining semantic fidelity. Empirical results further reveal a fundamental tradeoff in LLM-based metaphor translation: reducing figurative language often comes at the expense of input–output similarity. By introducing a controllable metaphor reduction rate parameter, our approach allows this balance to be tuned according to task requirements. Taken together, these innovations yield significant improvements over the state of the art.

[1] `https://www.ontologyportal.org`

2 Background

Research in metaphor handling divides into two complementary tasks: *detection*, which classifies text spans as metaphorical or literal, and *translation*, which transforms metaphorical expressions into literal equivalents that preserve meaning. This section reviews theoretical foundations, datasets, and computational approaches to both tasks.

2.1 Theoretical Foundations

Several linguistic theories inform computational approaches. The Metaphor Identification Procedure (MIP), developed by the Pragglejaz Group [11], provides a stepwise annotation framework for identifying linguistic metaphors. Steen et al. [12] extend it into Metaphor Identification Procedure VU University Amsterdam (MIPVU), achieving high inter-annotator agreement when building the Vrije University Amsterdam Metaphor Corpus (VUAMC). conceptual metaphor theory (CMT), proposed by Lakoff and Johnson [13], frames metaphor as mapping abstract domains (e.g., *time*) to concrete ones (e.g., *money*). Johnson's image-schematic conceptual metaphor (ISCM) [14] highlights the role of recurring sensorimotor patterns in shaping abstract reasoning. Finally, Wilks' selectional preference violation (SPV) theory [15, 16] detects metaphor via semantic mismatches between predicates and their arguments, a principle readily modeled using embeddings.

2.2 Datasets

Annotated corpora are central to training and evaluation. The VUAMC [12] remains the most widely used benchmark, covering four registers of the British National Corpus Baby Edition (BNC-Baby) [17] with 187k tokens and 13% metaphor density. MOH-X [18] provides verb-centered sentences labeled via crowdsourcing, while the TroFi dataset [19, 20] clusters literal and figurative verb usages from newswire. Other contributions include Bizzoni and Lappin's paraphrase-ranked corpus [21] and Stowe et al.'s Idiomatic and Metaphoric Paired Language Inference (IMPLI) dataset [22], which pairs figurative sentences with literal counterparts, offering rare resources for metaphor-to-literal translation.

2.3 Metaphor Detection

Supervised Approaches Deep learning has dominated detection research. Su et al.'s DeepNet [23] modeled detection as reading comprehension over token-level features. Choi et al.'s MelBERT [24] combined contextual and isolated embeddings to operationalize MIP and SPV, achieving state-of-the-art results on VUAMC and MOH-X. Zhang and Liu's MisNet [25] introduced dictionary-grounded embeddings to capture "basic meanings" and demonstrates improvements in domain-specific performance. Their follow-on work with adversarial

multitask learning (AdMul) [26] leveraged basic sense discrimination (BSD) data to enhance generalization. Other efforts, such as Wachowiak et al. [10], targeted ISCMs, detecting schema-specific metaphors with fine-tuned transformer models.

Large Language Models Recent work applies LLMs directly, mitigating data scarcity. Tian et al. [27] guided an LLM with knowledge graphs derived from metaphor theories, showing improved detection accuracy. Lin et al. [28] proposed a dual-perspective framework combining implicit retrieval-based guidance with explicit theory-driven prompting, leveraging both MIP and SPV representations to improve LLM performance. These approaches demonstrate that explicit integration of linguistic theory is crucial for effective zero-shot metaphor reasoning.

2.4 Metaphor Translation

Compared to detection, translation remains underexplored. Shutova [29] pioneered automatic paraphrasing of verb metaphors using distributional similarity and WordNet filtering, achieving strong human-judged accuracy. Mao et al. [30] advanced an unsupervised CBOW-based framework that simultaneously identified and paraphrased metaphors, demonstrating improved downstream machine translation performance. More recently, datasets like IMPLI [22] have enabled evaluation of literalization quality, though systematic models remain scarce.

2.5 Summary

Overall, research has established robust theories and datasets for metaphor detection, with transformer-based models achieving strong performance. However, translation lags behind, with only a handful of systems addressing the systematic rephrasing of metaphor into literal language. This gap motivates the present work, which investigates scalable metaphor resolution methods as a critical step toward trustworthy natural language to logic translation.

3 Methodology

This section describes the proposed metaphor resolution approach: *Batch Word-level Re-detection and Similarity Monitoring* (BWL). The method integrates automatic metaphor detection with large language model (LLM) prompting, iterative candidate generation, and prioritized evaluation. The pipeline operates at the token level to improve precision in metaphor elimination while balancing semantic fidelity.

3.1 Metaphor Detection

Metaphor detection was performed using Wachowiak et al.'s XLM-RoBERTa model [10], chosen for its robustness and reproducibility. The detector provided both sentence-level and token-level binary flags, allowing fine-grained identification of metaphorical words. Token-level detection was necessary to avoid over-reliance on strict sentence-level classification, which can fail when a candidate translation partially reduces metaphorical content.

3.2 Batch Candidate Generation

The BWL approach generated multiple translation candidates per prompt using an LLM (Llama 3.1, 8B parameters). Each prompt included both the input sentence and a list of metaphorically flagged words, which guided the model in selectively rephrasing figurative expressions. By explicitly providing this context, the LLM was relieved from independently identifying metaphors, improving translation consistency.

Figure 1 illustrates the overall pipeline. For a sentence S with M_S flagged words, a reduction rate parameter (RR) determined the maximum allowable metaphors (M_c) in any accepted candidate:

$$M_c = \lfloor M_S \, (1 - RR) \rfloor \tag{1}$$

The RR parameter, which ranges between zero and one, specifies the target proportion of metaphorical words to eliminate relative to the reference sentence. It effectively serves as a tunable threshold that balances the tradeoff between semantic similarity and metaphor reduction. At $RR = 1$, only candidates with zero residual metaphors are considered acceptable; however, if no such candidates exist, the architecture defaults to ranking purely on similarity. In practice, setting RR slightly below one (e.g., $RR = 0.9$) introduces flexibility, allowing translations with a small number of residual metaphors to remain in the highest priority bin (discussed in Section 3.3) when their similarity scores are strong, while still favoring those with fewer metaphors overall.

Batches of n candidates were generated per query, with model temperature optionally increased between iterations to promote lexical diversity. Each candidate was then re-analyzed by the detector to count remaining flagged words m_i and scored for semantic similarity (ss_i) using metrics such as Sentence-BERT (SBERT) cosine similarity [31], Bilingual Evaluation Understudy (BLEU) [32], and ROUGE-L [33].

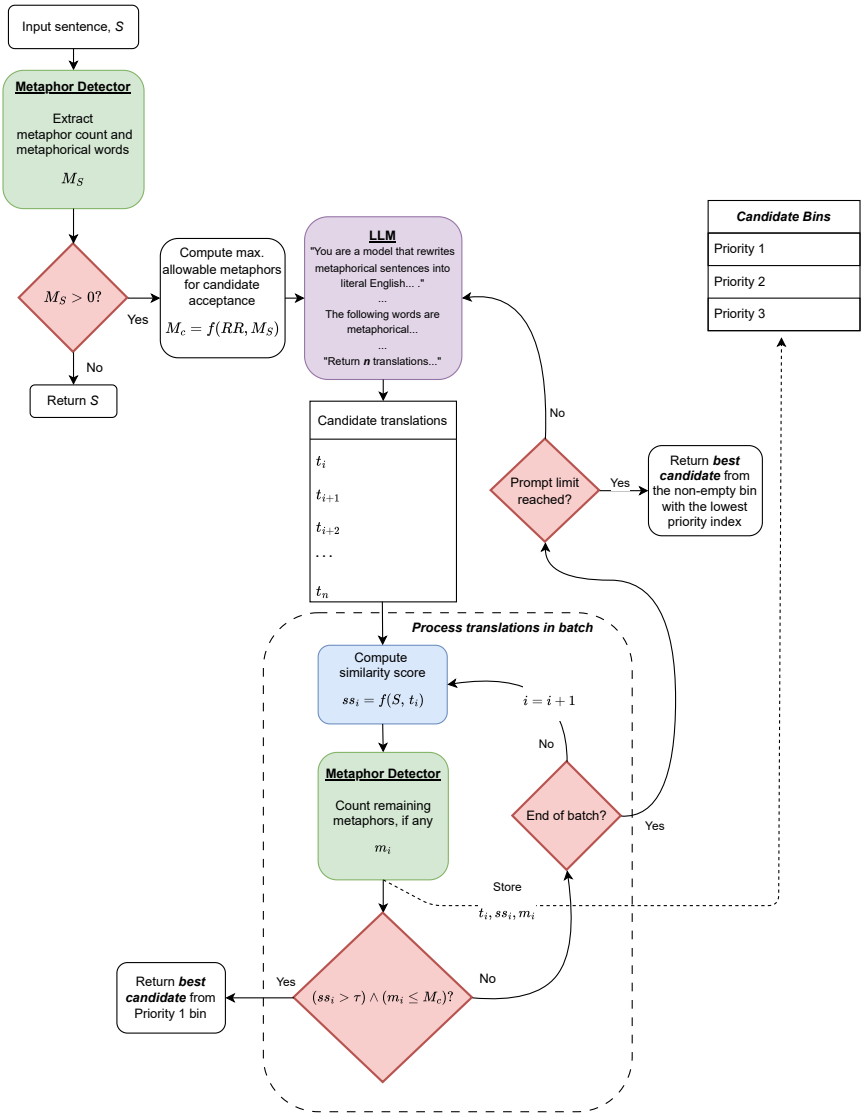

Fig. 1. batch word-level re-detection and similarity monitoring (BWL) architecture. Upon detecting M_S metaphorical words in the input sentence S, M_c was calculated as a function of M_S and RR. Candidate batches were generated and filtered until similarity and re-detection criteria were satisfied.

3.3 Candidate Prioritization

Each candidate t_i was placed into one of three bins based on the remaining metaphor count m_i and similarity score ss_i, as illustrated in Figure 2. Here, M_c is maximum allowable metaphorical words as defined in Equation 1 and τ denotes the user-defined similarity threshold. For example, if ROUGE is used as the similarity metric, a candidate must achieve a ROUGE score greater than τ, and a metaphor count less than M_c to be considered acceptable.

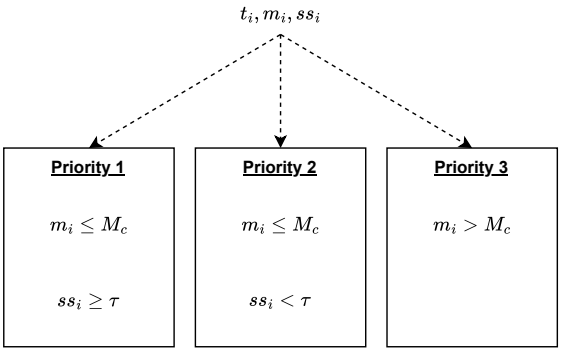

Fig. 2. Candidate translations were sorted into priority bins based on similarity and metaphor count. Selection favored minimum metaphors first, then maximum similarity.

During selection, bins were searched in priority order, with the highest-scoring candidate from the first non-empty bin returned as output. This mechanism explicitly balanced metaphor elimination against semantic fidelity, controlled by RR and τ.

3.4 Evaluation Setup

Experiments used the IMPLI dataset [22], reduced to 280 unique metaphor–literal pairs with punctuation corrected for JSON compatibility. Candidates were evaluated against reference literals using automated similarity metrics, while metaphor reduction was measured via re-detection. Equation 1 enabled flexible tuning: $RR = 0$ prioritized similarity only, while $RR = 1$ enforced complete elimination of flagged metaphors. Intermediate values balanced both criteria.

3.5 Summary

The BWL methodology combines token-level detection, batch candidate generation, flagged word injection, and prioritized evaluation. Figures 1 and 2 highlight its iterative filtering and binning design. By tuning RR and τ, the system

adapts to varying requirements for metaphor elimination and semantic preservation, providing a scalable approach to metaphor resolution for natural language to logic translation.

4 Results

This section reports results for the batch word-level re-detection and similarity monitoring (BWL) translation solution, using a zero-shot approach as a baseline for comparison. All experiments used the IMPLI dataset. Prior to testing, BWL parameters were optimized via a grid search. The zero-shot baseline is a single one-pass translation (no feedback, no iterated refinement).

4.1 Scoring and Composite Metric

We evaluate literalization quality by averaging the following metrics: BLEU, Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and SBERT cosine similarity, and metaphor removal with percent corrected and mean reduction ratio. By combining similarity measures with residual metaphor statistics, the composite metric unifies the two central evaluation criteria: faithfulness to the source text and effectiveness of metaphor removal, into a single score. For these similarity-based metrics, the gold-standard literal sentence served as the reference and the model's translation as the candidate. All reported values are averaged over the entire IMPLI dataset. The overall composite score C (used for model selection and headline comparison) was defined and weighted as follows:

$$C \;=\; w_1\,\overline{\mathrm{BERT}} \;+\; w_2\,\overline{\mathrm{BLEU}} \;+\; w_3\,\overline{\mathrm{ROUGE}} \;+\; w_4\,P_{\mathrm{corr}} \;+\; w_5\,\overline{RR}$$

$$w_1 = 0.15, \quad w_2 = 0.15, \quad w_3 = 0.1, \quad w_4 = 0.5, \quad w_5 = 0.1$$

where P_{corr} is the percent of translations with complete metaphor elimination, and $\overline{RR}$ is the average reduction ratio. The reduction ratio RR is

$$RR = \frac{M_{\mathrm{orig}} - M_{\mathrm{trans}}}{M_{\mathrm{orig}}} \tag{2}$$

where M_{orig} and M_{trans} are the numbers of flagged metaphorical words in the original and translated sentences, respectively.

4.2 Parameter Optimization

For BWL, the best grid point used `sim_function=bleu`, `des_sim_score=0.1`, `prompt_limit=4`, `batch_size=10`, and `red_rate=0.7`, with composite 0.660. A finer sweep of `red_rate` improved the composite to 0.679 at `red_rate=0.82` (from 0.660 $\rightarrow$ 0.679).

Metric	Zero-shot	BWL
Avg. BLEU	0.278	0.204
Avg. ROUGE	0.481	0.390
Avg. SBERT Cosine	0.906	0.879
Percent Corrected	0.271	0.775
Avg. Reduction Ratio	0.315	0.903
Composite	**0.393**	**0.679**

Table 1. Overall performance summary of optimized translation approaches (zero-shot vs. BWL)

4.3 Aggregate Metrics: Zero-Shot vs. BWL

Table 1 summarizes the post-optimization results used in the final comparison. While zero-shot attains higher similarity scores (BLEU/ROUGE/SBERT), BWL achieves markedly superior metaphor removal (percent corrected +50.4 points; reduction ratio +0.588), yielding a much higher composite (+0.286 absolute). The inclusion of originally flagged words in the BWL prompts, larger candidate pools (via batches), and the reduction-rate criterion together drive these gains.

4.4 Qualitative Illustration

Below we reproduce the required BWL translation example (verbatim fragment emphasized):

Original:
Four of the absentees suffered the squirming discomfort of being among the Welsh squad.
Detected metaphorical words: {squirming, discomfort, squad}

Gold Reference:
Four of the absentees suffered the continuous discomfort of being among the Welsh squad.
Detected metaphorical words: {discomfort, squad}

BWL Translation:
The four absentees didn't like being among the group of Welsh players.
Remaining metaphorical words: {}
BLEU score (with respect to gold reference): **0.142**

The BWL output removes metaphorical constructions while preserving propositional content, illustrating the broader trade-off we observed: stronger literalization often requires departing from the exact surface form of the gold reference. Importantly, the detector still flagged metaphors in the gold reference, even though the IMPLI annotations indicated only a single metaphor ("squirming") in this example. This mismatch could stem from dataset limitations or

detector false positives, which in turn forced the translation system to remove words that were not truly metaphorical.

4.5 Metaphor Elimination Outcome Categories

To analyze elimination behavior at the sentence level, we partition translations into outcome categories (summarized here): *false negatives* (detected as non-metaphorical initially), *pure elimination* (all original metaphors removed, none added), *pure reduction* (strict subset of original metaphors retained), *impure reduction* (fewer metaphors overall but at least one new metaphor introduced), *full replacement* (same count as original but all metaphors replaced by different ones), *partial replacement* (same count with some originals retained), *inflation* (more metaphors than the original), and *no change*. Selected examples of original–translation sentence pairs representative of each category are presented in Table 2.

Category	**Original / Translated Sentence**	**Flagged Met. Words**
False Negative	**Original:** The plane climbs reluctantly, one set of wings dipping drunkenly. **Translation:** The plane climbs reluctantly, one set of wings dipping drunkenly.	Original: [] Translation: []
Pure Elimination	**Original:** There are few things worse than being bludgeoned into reading a book you hate. **Translation:** There are few bad experiences than being forced to read a book you don't like.	Original: [things, bludgeoned, into] Translation: []
Pure Reduction	**Original:** The ostentatious way of living of the rich ignites the hatred of the poor. **Translation:** The rich people's luxurious way of living is not liked by the poor.	Original: [way, ignites] Translation: [way]
Impure Reduction	**Original:** The steering of my new car answers to the slightest touch. **Translation:** The steering of my new car responds when you touch it lightly.	Original: [answers, to] Translation: [responds]
Partial Replacement	**Original:** It dawned on him that she had betrayed him. **Translation:** It landed on him that she had betrayed him.	Original: [dawned, on] Translation: [landed, on]
Full Replacement	**Original:** The new measures deflated the economy. **Translation:** The new rules hurt the economy.	Original: [measures, deflated] Translation: [rules, hurt]
Inflation	**Original:** Sales were climbing after prices were lowered. **Translation:** Sales went up after prices were lowered.	Original: [climbing] Translation: [went, up, lowered]
No Change	**Original:** The bad review of his work deflated his self-confidence. **Translation:** The negative review of his employment deflated his self-confidence.	Original: [deflated] Translation: [deflated]

Table 2. Examples of metaphor translation outcomes by category, with flagged metaphorical words underlined. Note that the example sentence in the false negative category contains at least one obvious metaphor, yet none were flagged.

Table 3 presents the category statistics for zero-shot and BWL only (counts and percentages taken verbatim from the full analysis).

Observations. BWL converts a large majority of cases to *pure elimination* (77.5%, vs.27.1% for zero-shot) and sharply reduces undesirable outcomes that add or swap metaphors (*impure reduction*, *full replacement*, *inflation*). Together

	Zero-shot		BWL	
Category	Count	%	Count	%
False negatives	38	13.6%	**38**	**13.6%**
Pure elimination	76	27.1%	**217**	**77.5%**
Pure reduction	11	3.9%	**2**	**0.7%**
Impure reduction	57	20.4%	**10**	**3.6%**
Full replacement	62	22.1%	**7**	**2.5%**
Partial replacement	0	0.0%	**0**	**0.0%**
Inflation	36	12.9%	**6**	**2.1%**
No change	0	0.0%	**0**	**0.0%**

Table 3. Metaphor elimination outcomes across translation solutions (counts and percentages)

with the much higher reduction ratio, this indicates that BWL reliably removes metaphorical content without introducing new figurative language.

4.6 Summary of Improvements over Zero-Shot

Quantitatively, BWL improves percent corrected from 0.271 → 0.775 and average reduction ratio from 0.315 → 0.903, raising the composite from 0.393 → 0.679. Qualitatively, BWL produces cleaner literalizations (e.g., "*the* ***four absentees didn't like*** *being among the group of Welsh players*") with far fewer cases of metaphor introduction or substitution. While similarity scores are modestly lower than zero-shot—reflecting fewer surface-level matches to gold references—the overall objective of metaphor elimination is achieved substantially more often with BWL.

5 Conclusions

This work developed and evaluated a metaphor resolution framework for use in autoformalization, emphasizing the translation of metaphorical English sentences into literal paraphrases. The proposed BWL approach leverages an LLM to generate batches of diverse candidate translations, which are ranked according to similarity and re-detection criteria. By selecting the best candidate, the method achieved a substantial improvement over a simple zero-shot baseline, with 50% higher rates of complete metaphor removal and 59% higher rates of metaphor reduction. The optimized BWL configuration attained a composite score of 0.679, compared to 0.393 for zero-shot, and consistently outperformed in metaphor elimination categories while preserving adequate semantic similarity.

The findings underscore the importance of a dedicated metaphor detector within the translation pipeline. Transformer-based detection models proved essential not only for initial identification but also for validating candidate translations. Analysis further reveals that while LLMs are strong paraphrasers, they

frequently fail to recognize or avoid metaphors, often introducing new figurative expressions in zero-shot or single-response prompting. Batch-based methods, in contrast, encourage more diverse and creative paraphrases that more effectively eliminate metaphorical content.

Several limitations emerged. The tradeoff between semantic similarity and metaphor elimination remains significant: translations that achieve near-total elimination tend to deviate more from the original sentence structure. Additionally, the evaluation relied on automatic metrics and assumed an oracle-like detector, whereas real-world detectors may introduce false positives or negatives that propagate through the pipeline. Finally, the reliance on a single open-source LLM constrains generalizability across models and domains.

Future work should explore advanced prompting strategies, such as chain-of-thought prompting, to combine improved similarity with robust elimination. Incorporating LLM orchestration frameworks would also enable more flexible and maintainable implementations of complex prompting strategies. For HyNOLU applications, integrating word sense disambiguation (WSD) could bypass unnecessary translation of metaphors already mapped to SUMO, streamlining processing. Broader experimentation with alternative LLMs and further refinement of metaphor detectors, particularly models with high F-1 performance, represent critical directions to improve translation quality and reliability.

In summary, this research demonstrates that effective metaphor resolution can be achieved by combining a specialized detection model with batch-based LLM translation. The BWL approach provides a practical and scalable solution, advancing the capability of the HyNOLU pipeline to produce literal, logic-ready language suitable for automated reasoning tasks.

References

[1] Yue Zhang et al. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. arXiv. 2025. DOI: 10.48550/arXiv.2309.01219.

[2] Parshin Shojaee et al. *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. arXiv. 2025. DOI: 10.48550/arXiv.2506.06941.

[3] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. arXiv. 2017. DOI: 10.48550/arXiv.1708.08296. eprint: 1708.08296.

[4] Richard Thompson et al. "Formalizing Natural Language: Cultivating LLM Translations Using Automated Theorem Proving". In: *Theorem Proving and Machine Learning in the Age of LLMs: State of the Art and Future Perspectives*. EuroProofNet, COST Action CA20111 – European Research Network on Formal Proofs. Edinburgh, Scotland, UK, Apr. 2025. URL: https://europroofnet.github.io/wg5-edinburgh25/.

[5] Adam Pease. *SUO-KIF Reference Manual*. https://github.com/ontologyportal/sigmakee/blob/master/suo-kif.pdf. retrieved 20 June 2025. 2009.

[6] Ian Niles and Adam Pease. "Toward a Standard Upper Ontology". In: *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*. Ed. by Chris Welty and Barry Smith. 2001, pp. 2–9.

[7] Adam Pease. *Ontology: A Practical Guide*. Angwin, CA: Articulate Software Press, 2011.

[8] Stephan Schulz. "E–a brainiac theorem prover". In: *AI Communications* 15.2-3 (2002). Available: ResearchGate, pp. 111–126.

[9] Laura Kovács and Andrei Voronkov. "First-order theorem proving and Vampire". In: *International Conference on Computer Aided Verification*. Springer. 2013, pp. 1–35.

[10] Lennart Wachowiak, Dagmar Gromann, and Chao Xu. "Drum Up SUPPORT: Systematic Analysis of Image-Schematic Conceptual Metaphors". In: *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. 2022. DOI: 10.18653/v1/2022.flp-1.7.

[11] Pragglejaz Group. "MIP: A Method for Identifying Metaphorically Used Words in Discourse". In: *Metaphor and Symbol* 22.1 (2007), pp. 1–39. DOI: 10.1080/10926480709336752.

[12] Gerard J. Steen et al. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam, Netherlands: John Benjamins Publishing Company, 2010.

[13] George Lakoff and Mark Johnson. *Metaphors We Live By*. Chicago, IL, USA: University of Chicago Press, 1980.

[14] Mark Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Reason, and Imagination*. Chicago, IL, USA: University of Chicago Press, 1987.

[15] Yorick Wilks. "A Preferential, Pattern-Seeking, Semantics for Natural Language Inference". In: *Words and Intelligence I: Selected Papers by Yorick Wilks*. Ed. by Khurshid Ahmad, Christopher Brewster, and Mark Stevenson. Dordrecht, Netherlands: Springer Netherlands, 2007, pp. 83–102. ISBN: 978-1-4020-5285-9. DOI: 10.1007/1-4020-5285-5_5.

[16] Yorick Wilks. "Making preferences more active". In: *Words and Intelligence I: Selected Papers by Yorick Wilks*. Ed. by Khurshid Ahmad, Christopher Brewster, and Mark Stevenson. Dordrecht, Netherlands: Springer Netherlands, 2007, pp. 141–166. DOI: 10.1007/1-4020-5285-5_7.

[17] British National Corpus Consortium. *British National Corpus 1994, Baby edition*. Literary and Linguistic Data Service, 2007. URL: http://hdl.handle.net/20.500.14106/2553.

[18] Saif Mohammad, Ekaterina Shutova, and Peter Turney. "Metaphor as a Medium for Emotion: An Empirical Study". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 2016. DOI: 10.18653/v1/S16-2003. URL: https://aclanthology.org/S16-2003/.

[19] Julia Birke and Anoop Sarkar. "A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Available: ACL Anthology. 2006.

[20] Julia Birke and Anoop Sarkar. "Active Learning for the Identification of Nonliteral Language". In: *Proceedings of the Workshop on Computational Approaches to Figurative Language*. Available: ACL Anthology. 2007.

[21] Yuri Bizzoni and Shalom Lappin. "Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks". In: *Proceedings of the Workshop on Figurative Language Processing*. 2018. DOI: 10.18653/v1/W18-0906.

[22] Kevin Stowe, Prasetya Utama, and Iryna Gurevych. "IMPLI: Investigating NLI Models' Performance on Figurative Language". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022. DOI: 10.18653/v1/2022.acl-long.369.

[23] Chuandong Su et al. "DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection". In: *Proceedings of the Second Workshop on Figurative Language Processing*. 2020. DOI: 10.18653/v1/2020.figlang-1.4. (Visited on 01/10/2025).

[24] Minjin Choi et al. *MelBERT: Metaphor Detection via Contextualized Late Interaction Using Metaphorical Identification Theories*. arXiv. 2021. DOI: 10.48550/arXiv.2104.13615.

[25] Shenglong Zhang and Ying Liu. "Metaphor Detection via Linguistics Enhanced Siamese Network". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Available: ACL Anthology. 2022.

[26] Shenglong Zhang and Ying Liu. *Adversarial Multi-Task Learning for End-to-End Metaphor Detection*. arXiv. 2023. DOI: 10.48550/arXiv.2305.16638.

[27] Yuan Tian, Nan Xu, and Wenji Mao. "A Theory Guided Scaffolding Instruction Framework for LLM-Enabled Metaphor Reasoning". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2024. DOI: 10.18653/v1/2024.naacl-long.428.

[28] Yujie Lin et al. *A Dual-Perspective Metaphor Detection Framework Using Large Language Models*. arXiv. 2024. DOI: 10.48550/arXiv.2412.17332.

[29] Ekaterina Shutova. "Automatic Metaphor Interpretation as a Paraphrasing Task". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Available: ACL Anthology.

[30] Rui Mao, Chenghua Lin, and Frank Guerin. "Word Embedding and WordNet Based Metaphor Identification and Interpretation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018. DOI: 10.18653/v1/P18-1113.

[31] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks*. arXiv. 2019. DOI: 10.48550/arXiv.1908.10084.

[32] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002. DOI: 10.3115/1073083.1073135.

[33] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Available: ACL Anthology. 2004.