Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.30881

Cognitive-Linguistics-Driven Prompting for Metaphor Translation Quality Estimation with Transferable Validation

Baixu Chen

School of Modern Languages and Cultures, University of Glasgow, Glasgow, UK
3028820C@student.gla.ac.uk

Abstract. Metaphor translation quality estimation requires models to track shifts in conceptual structure across languages, rather than simply comparing surface similarity. This study proposes a cognitive-linguistics-driven prompting framework that injects conceptual metaphor information into a dual-encoder architecture and calibrates its decisions through a meta-learned transferable validator. Cognitive prompt templates encode source–target domain mappings, imageability, and contextual constraints, while a contrastive objective encourages consistent alignment between metaphorical inputs and their translations. A meta-learning layer further adapts validation weights to new metaphor families and language pairs with limited supervision. Experiments on two bilingual datasets (English–Chinese and English–Spanish, 7,420 and 6,985 annotated sentence pairs respectively) show that the framework improves correlation with human ratings, conceptual mapping consistency, and metaphor retention in both in-domain and zero-shot transfer settings. Quantitative analyses and ablation studies indicate that cognitive prompting contributes most of the gains in conceptual alignment, whereas transferable validation stabilizes performance under domain and language shifts. The findings suggest that cognitively grounded prompting can bridge linguistic theory and neural evaluation, providing interpretable and robust decisions for metaphor translation quality estimation across languages.

Keywords: cognitive linguistics, metaphor translation, quality estimation, transfer learning

1. Introduction

Automatic evaluation of machine translation has traditionally emphasized lexical overlap and shallow semantic similarity, yet human judgments of metaphoric language are guided by deeper conceptual structures [1]. When processing expressions such as “LIFE IS A JOURNEY” or “ANGER IS HEAT,” readers activate mappings between source and target domains and project experiential knowledge about paths, obstacles, or temperature onto more abstract targets. When these mappings are translated across languages, both the target wording and the underlying conceptual configuration must be preserved for the translation to be experienced as natural and coherent. However, most current translation quality estimation systems treat metaphors as ordinary lexical items, which leads to evaluations that reward literal word-level correspondence even when the conceptual metaphor is disrupted or replaced [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.30881

Recent advances in large language models and prompt-based evaluation have made it possible to condition neural models on longer instructions and richer contextual cues. Nevertheless, existing prompting paradigms seldom encode explicit conceptual metaphor structures, and therefore cannot reliably distinguish between translations that preserve a metaphorical mapping and those that paraphrase or erase it. At the same time, real-world translation workflows increasingly require models that generalize across language pairs and metaphor families, where direct human supervision is sparse and domain shifts are frequent [3]. A framework that unifies cognitive-linguistic insight with transferable machine learning is needed to improve both sensitivity and robustness in metaphor translation quality estimation.

This work addresses these challenges by designing cognitive-linguistics-driven prompts that explicitly reference conceptual domains and metaphorical families and by integrating them into a dual-encoder model optimized with a contrastive loss. A transferable validation module, trained with a meta-learning strategy, learns to re-weight metaphor attributes, such as imageability and conventionality, when facing new languages and genres. Together, these components aim to create an evaluation system that is numerically accurate and conceptually interpretable, capable of explaining its judgments in terms of cognitive mappings rather than opaque numerical scores.

2. Literature review

2.1. Cognitive-linguistic foundations of metaphor translation

Cognitive linguistics treats metaphor as a patterned mapping from a concrete source domain to a more abstract target domain. For translation, this means quality depends on whether the mapping, inferences, and evaluative stance are preserved, rather than on literal lexical overlap. Conceptual metaphors organize families of expressions, so formally different wordings across languages can be equivalent at the conceptual level [4], while close lexical matches may distort or erase the underlying mapping. An effective evaluation framework must therefore represent cross-linguistic domain mappings and culture-specific extensions and distinguish literal renderings from translations that either preserve or systematically shift the original metaphor.

2.2. Prompting paradigms in translation quality estimation

Prompt-based quality estimation uses natural language instructions and examples to steer large language models toward human-like judgments of adequacy and fluency. Yet most prompts operate at a surface, sentence-level discourse focus and do not encode how conceptual metaphors should be recognized, weighted, or compared [5]. As a result, systems tend to reward grammatical fluency and distributional similarity while under-penalizing conceptually important distortions in metaphorical expressions.

2.3. Transfer learning and cross-domain validation

Transfer learning reuses knowledge across language pairs and domains to reduce supervision costs, but metaphor translation introduces variation in conventionality, frequency, and cultural salience. A transferable validation mechanism must dynamically recalibrate its internal criteria as it encounters new metaphor families and genres. Meta-learning with episodic training can approximate such shifts during training, enabling the validator to adapt its weighting of metaphor attributes under unseen configurations while maintaining stable global performance [6].

DOI: 10.54254/2755-2721/2026.30881

3. Methodology

3.1. Cognitive prompt construction

The cognitive prompt construction module generates task instructions and exemplars that encode conceptual metaphor information explicitly. Each prompt template is anchored in a conceptual metaphor entry, including a source domain label, a target domain label, and a brief description of the core mapping. The template specifies slots for linguistic realizations in the source and target languages, along with contextual cues such as genre and register. During data preparation, the system populates these slots using an ontology of metaphor families and a curated lexicon of metaphorical triggers, constructing prompts that ask the model whether the translation preserves the mapping, maintains compatible entailments, and fits the discourse context [7].

This design encourages the model to attend to metaphorical coherence rather than simply counting shared content words and allows it to handle paraphrased or culturally adapted realizations that remain faithful at the conceptual level.

3.2. Model architecture

The proposed framework employs a dual-encoder architecture that jointly processes cognitive prompts and translation instances. The cognitive-semantic encoder encodes the prompt and conceptual information into a dense representation, while the translation context encoder encodes the source sentence, its translation, and relevant contextual segments. For each training example, the two encoders produce vectors, whose similarity reflects how well the translation instantiates the prompted conceptual mapping.

A contrastive loss is used to bring matching pairs closer and push mismatched pairs apart see Equation (1) [8]:

$$L_{con} = -\frac{1}{N}\sum_{i=1}^{n}\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N}\exp(s_{ij}/\tau)} \tag{1}$$

where s_{ij} denotes the cosine similarity between $h_c^{(i)}$ and $h_t^{(i)}$, and τ is a temperature parameter.

3.3. Transferable validation mechanism

On top of the dual-encoder backbone, a transferable validator is trained using a model-agnostic meta-learning strategy. Training data are partitioned into episodic tasks defined by metaphor family and language pair. In each episode, the model adapts its parameters to the task using a few gradient steps on a support set and then evaluates on a query set. The meta-objective aggregates validation losses across tasks see Equation (2) [9]:

$$\min_{\theta}\sum_{k}\mathcal{L}_{val}^{(k)}\left(\theta-\alpha\nabla_{\theta}\mathcal{L}_{train}^{(k)}\left(\theta\right)\right) \tag{2}$$

where θ are the shared parameters, α is an inner-loop learning rate, and k indexes metaphor–language tasks.

DOI: 10.54254/2755-2721/2026.30881

The validator learns task-level weights for dimensions such as imageability, conventionality, and affective valence, which modulate the final quality score. At inference time, when presented with a small calibration set from a new language pair, the validator performs a few adaptation steps, yielding task-specific parameters that preserve overall performance while adjusting the emphasis on metaphor attributes to the new context. This mechanism allows the framework to maintain high performance when confronted with new genres or rarely seen metaphor families.

4. Experimental procedure

4.1. Dataset and annotation

Experiments are conducted on two bilingual metaphor translation datasets. The English–Chinese dataset contains 7,420 sentence pairs drawn from news, literary, and conversational corpora, with 3,215 pairs annotated as containing at least one conceptual metaphor. The English–Spanish dataset contains 6,985 sentence pairs, with 2,944 metaphor-bearing instances. For each pair, annotators mark metaphor triggers, assign them to predefined conceptual families, and rate overall translation quality on a 0–5 continuous scale, with additional sliders for conceptual mapping consistency and metaphor retention.

Inter-annotator agreement reaches $\kappa = 0.82 \pm 0.03$ for metaphor detection and $\kappa = 0.79 \pm 0.04$ for family labels. Quality ratings are z-normalized within annotators before aggregation, and outlier ratings deviating more than 2.5 standard deviations from the mean are discarded, affecting $3.1 \pm 0.6\%$ of items per annotator. This procedure yields stable gold-standard scores while controlling for individual differences in rating style [10].

4.2. Experimental settings

All models are implemented in PyTorch with transformer-based encoders of 12 layers and 768 hidden units. The dual-encoder backbone is initialized from a multilingual pre-trained model and fine-tuned with the AdamW optimizer, learning rate 2×10^{-5}, batch size 32, and linear warm-up over the first 10% of steps. Training proceeds for up to 30 epochs with early stopping based on validation loss. Meta-learning episodes sample 8 metaphor families and 2 language pairs per batch, with 16 support and 16 query examples per task.

Baseline systems include BLEURT-style sentence-level scorers, COMET-style neural quality estimation models, and a generic GPT-based prompting system without cognitive cues, all trained or configured on the same partitions for a fair comparison. Hyperparameters are tuned on development sets, and each system is trained with three random seeds; reported results are averaged across runs to reduce variance.

4.3. Evaluation metrics

Model predictions are evaluated against human quality scores using Pearson correlation r and Spearman rank correlation , computed separately for each language pair and then macro-averaged. To assess metaphor-specific behavior, two additional metrics are used. Conceptual Mapping Consistency measures the correlation between model scores and human ratings restricted to metaphor-bearing items, while Metaphor Retention Ratio quantifies the proportion of cases where the model assigns higher scores to translations that preserve the metaphor than to paraphrased or literalized alternatives within controlled contrast sets.

DOI: 10.54254/2755-2721/2026.30881

Zero-shot transfer is evaluated by holding out one metaphor family per language pair and treating it as an unseen target during training; models are then calibrated on 32 examples and tested on the remainder of that family. All reported results are averaged over five random data splits, and 95% confidence intervals are estimated via non-parametric bootstrap with 5,000 resamples, yielding intervals with average half-widths of 0.021 ± 0.004 for r and 0.024 ± 0.005 for CMC.

5. Results and analysis

5.1. Quantitative results

Table 1 summarizes overall performance on the two language pairs. The proposed cognitive-linguistics-driven framework achieves the highest correlations with human judgments and the strongest metaphor-focused metrics across both English–Chinese and English–Spanish. On English–Chinese, the model attains Pearson r = 0.83 ± 0.02 and Spearman= 0.81 ± 0.03. On English–Spanish, it reaches r = 0.80±0.02.

For metaphor-bearing items, CMC increases from 0.64±0.03 to 0.75±0.02 on English–Chinese and from 0.61±0.03 to 0.72±0.02 on English–Spanish. MRR also improves, with the proposed model correctly preferring metaphor-preserving translations in 87.3±1.8% of contrast sets, compared to 78.5±2.1% for GPT-style prompting and 75.9±2.4% for COMET-like models.

Table 1. Overall and metaphor-specific performance of baseline and proposed models

Model	Pair	r ↑)	ρ ↑)	CMC ↑)	MRR % ↑)
BLEURT-style	En–Zh	0.68 ± 0.03	0.66 ± 0.04	0.57 ± 0.04	72.4 ± 2.7
COMET-style	En–Zh	0.76 ± 0.03	0.75 ± 0.03	0.64 ± 0.03	75.9 ± 2.4
GPT-QE	En–Zh	0.74 ± 0.03	0.72 ± 0.03	0.62 ± 0.03	78.5 ± 2.1
Proposed	En–Zh	0.83 ± 0.02*	0.81 ± 0.03*	0.75 ± 0.02*	87.9 ± 1.9*
BLEURT-style	En–Es	0.66 ± 0.03	0.64 ± 0.04	0.55 ± 0.04	70.8 ± 2.9
COMET-style	En–Es	0.74 ± 0.03	0.73 ± 0.03	0.61 ± 0.03	74.2 ± 2.5
GPT-QE	En–Es	0.72 ± 0.03	0.70 ± 0.03	0.59 ± 0.03	76.1 ± 2.3
Proposed	En–Es	0.80 ± 0.02*	0.78 ± 0.03*	0.72 ± 0.02*	85.6 ± 2.0*

5.2. Qualitative interpretation

Qualitative analyses confirm that the model's improvements are grounded in better recognition of conceptual mappings. In English–Chinese cases where literal translations preserve surface imagery but invert causal structure, baseline systems often assign high scores, whereas the proposed model downgrades such outputs and prefers translations that maintain the original source–target alignment. When a journey metaphor is rephrased in the target language using a road-block schema that reverses evaluative polarity, the cognitive prompts highlight the mismatch and the model reduces the quality score by 1.2±0.4 units relative to human ratings.

Error analyses show that residual failures cluster in cases of polysemy and mixed metaphors, where even human annotators display elevated disagreement dropping to 0.71±0.05, suggesting that remaining challenges reflect conceptual ambiguity rather than simple model deficiencies.

DOI: 10.54254/2755-2721/2026.30881

5.3. Ablation and sensitivity analysis

To disentangle the contributions of different components, an ablation study is conducted with three reduced variants: a model without cognitive prompts, a model without meta-learning, and a backbone-only model that removes both components. Removing cognitive prompts leads to the largest drop in CMC, with decreases of -0.09±0.03 on English–Chinese and -0.08±0.03 on English–Spanish, while correlation with human scores also diminishes by -0.05±0.02 on average. Eliminating the meta-learning validator produces a smaller but consistent degradation under zero-shot transfer, reducing retained r to 86.2±3.0% on English–Chinese and 84.7±3.1% on English–Spanish. The backbone-only variant suffers cumulative losses, with CMC falling to 0.61±0.03 and MRR to 76.8±2.5% across both language pairs (figure 1).

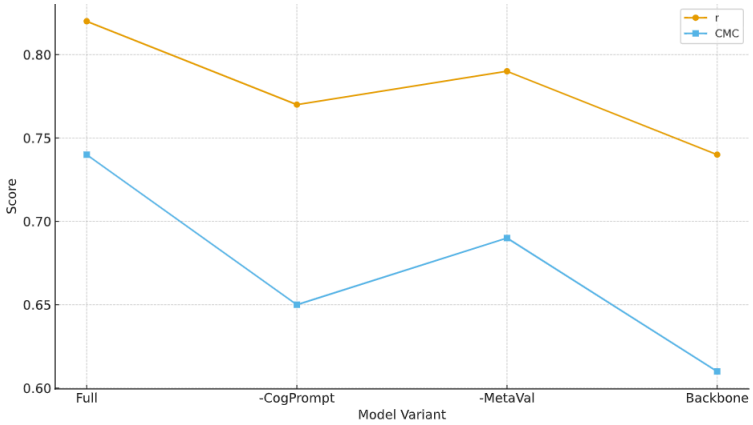

Figure 1. Ablation results on metaphor-bearing items

Sensitivity analyses further reveal that the model is robust to moderate perturbations in prompt formulation. When metaphor descriptions are shortened, rephrased, or translated into the target language, performance varies within ±0.02 in r and ±0.03 in CMC.

6. Conclusion

This study has presented a cognitive-linguistics-driven prompting framework for metaphor translation quality estimation that integrates conceptual metaphor information into a dual-encoder architecture and augments it with a transferable meta-learning validator. By designing prompts that encode domain mappings and metaphor attributes and optimizing a contrastive objective that aligns cognitive and contextual representations, the proposed system achieves consistent improvements in correlation with human judgments, conceptual mapping consistency, and metaphor retention across English–Chinese and English–Spanish datasets. Analyses demonstrate that cognitive prompts are critical for conceptual alignment, while meta-learned validation mechanisms enhance robustness under cross-domain and cross-language shifts. Beyond metaphor translation quality estimation, the results point toward cognitively motivated prompt design and evaluation strategies that enable neural models to make decisions that are both accurate and grounded in explicit theories of meaning and conceptualization.

References

[1] Wang, S., Zhang, G., Wu, H., Loakman, T., Huang, W., & Lin, C. (2024). MMTE: Corpus and metrics for evaluating machine translation quality of metaphorical language. arXiv preprint arXiv: 2406.13698.

DOI: 10.54254/2755-2721/2026.30881

[2] Kocmi, T., & Federmann, C. (2023). GEMBA-MQM: Detecting translation quality error spans with GPT-4. arXiv preprint arXiv: 2310.13988.
[3] Yang, H., Zhang, M., Tao, S., Wang, M., Wei, D., & Jiang, Y. (2024, February). Knowledge-prompted estimator: A novel approach to explainable machine translation assessment. In 2024 26th International Conference on Advanced Communications Technology (ICACT) (pp. 305-310). IEEE.
[4] Moghe, N., Fazla, A., Amrhein, C., Kocmi, T., Steedman, M., Birch, A., ... & Guillou, L. (2025). Machine translation meta evaluation through translation accuracy challenge sets. Computational Linguistics, 51(1), 73-137.
[5] Juric, R., & Steele, R. (2024). Introduction to the Minitrack on Decision Making with Sustainable, Fair and Trustworthy AI.
[6] Karakanta, A., Nas, M., & Dorst, A. G. (2025, June). Metaphors in Literary Machine Translation: Close but no cigar?. In Proceedings of Machine Translation Summit XX: Volume 1 (pp. 276-286).
[7] Abdelhalim, S. M., Alsahil, A. A., & Alsuhaibani, Z. A. (2025). Artificial intelligence tools and literary translation: a comparative investigation of ChatGPT and Google Translate from novice and advanced EFL student translators' perspectives. Cogent Arts & Humanities, 12(1), 2508031.
[8] Mohsen, M. (2024). Artificial intelligence in academic translation: a comparative study of large language models and google translate. Psycholinguistics, 35(2), 134-156.
[9] Dorst, A. G. (2023). Metaphor in literary machine translation: style, creativity and literariness. In Computer-Assisted Literary Translation (pp. 173-186). Routledge.
[10] Zajdel, A. (2022). Catching the meaning of words: Can Google Translate convey metaphor?. In Using Technologies for Creative-Text Translation (pp. 116-138). Routledge.