SABRE: Semantic Anchored Biaffine Relation Extraction for Chinese Metaphors

Mingchao Liu, and Shuo Wang*
Hebei Key Laboratory of Machine Learning and Computational Intelligence
College of Mathematics and Information Science
Hebei University
Baoding 071002, China
Email: 20231105036@stumail.hbu.edu.cn, shuowang@hbu.edu.cn

***Abstract*—Metaphorical Relation Extraction (MRE) is a fundamental task that bridges linguistic expressions and conceptual mappings by identifying directed connections between *Target* and *Source* spans. Current state-of-the-art methods typically rely on a *generate-then-classify* paradigm, where relations are predicted only for pre-filtered candidate spans. This cascaded structure introduces irreversible error propagation, resulting in a severe "Recall Bottleneck" in which implicit metaphors are discarded at early stages. In this paper, we propose SABRE, a unified end-to-end structured decoding framework that jointly optimizes span representation learning and metaphor relation extraction within a single inference stage, without cascaded pruning. Specifically, drawing inspiration from recent prompt-based extraction approaches, we construct a label-augmented input sequence and introduce Semantic Anchors to explicitly probe the pre-trained encoder for subtle metaphorical cues, enabling joint span representation learning without inference-stage pruning decisions. Furthermore, we employ a deep Biaffine mechanism to rigorously model the directional dependencies between spans. To address the severe class imbalance inherent in MRE, a Focal Loss objective is incorporated. Experimental results on the benchmark CMRE dataset demonstrate that SABRE achieves a new state-of-the-art F1 score of 67.04% using the standard BERT backbone, yielding a 3.58% absolute improvement over the baseline.**

***Index Terms*—Metaphorical Relation Extraction, Prompt Learning, Biaffine Attention, Natural Language Processing**

I. Introduction

Metaphor is widely used in human communication [1] and serves as a fundamental cognitive mechanism through which humans conceptualize abstract concepts (Tenor) in terms of concrete experiences (Vehicle) [2]. To bridge the gap between surface-level linguistic identification and deeper conceptual mapping, Chen et al. [3] formalized the task of *Chinese Metaphorical Relation Extraction* (MRE), which aims to identify directed semantic relations between metaphorically related spans. Unlike traditional token-level sequence labeling, MRE explicitly models asymmetric mappings, such as linking "*drinking*" (Source Action) to "*car*" (Target) in the sentence "*My car is drinking gas*."

Although recent studies have made progress in *metaphor generation* [4], accurate *extraction* remains the cornerstone for understanding complex metaphorical structures and enabling downstream semantic reasoning. The current state-of-the-art approach for MRE [3] is formulated as an end-to-end span-based joint model. However, its inference procedure follows a *generate-then-classify* paradigm, which introduces two fundamental limitations.

Cascaded Error Propagation. Although trained jointly, previous models decode relations by first generating and pruning candidate spans. This inference factorization introduces *structural irreversibility*: if a subtle or implicit metaphor candidate (e.g., an abstract attribute) is discarded during the initial span generation phase, the downstream relation module has *no opportunity* to recover it. This dependency creates a severe "Recall Bottleneck," particularly for metaphors that lack distinct surface features. Indeed, the error analysis in [3] reveals that nearly **60.0%** of errors are attributed to "Fail to Recall", highlighting a structural limitation of generate-then-classify decoding for metaphorical relation extraction.

Weak Directionality Modeling. Metaphorical mappings are inherently asymmetric and directional (Source $\rightarrow$ Target). However, existing methods typically adopt symmetric span-pair representations, such as vector concatenation followed by linear classification. Such decoders lack explicit inductive bias for directionality, often leading to reversed or spurious relations, especially in sentences containing complex metaphorical structures.

To address these challenges, we propose **SABRE** (**S**emantic **A**nchored **Bi**affine **R**elation **E**xtractor). Instead of factorizing inference into span generation followed by relation classification, SABRE jointly performs span representation learning and relation decoding within a unified structured prediction framework. This unified **one-stage** design inherently avoids the irreversible error propagation of cascaded models.

Specifically, drawing inspiration from the emerging **prompt-based extraction paradigm** [5], [6], we reformulate MRE as a *structure-aware semantic matching* task. Unlike traditional methods that rely on fixed classification heads, we introduce **Semantic Anchors**—label-aware tokens prepended to the input—to explicitly query the encoder. This mechanism allows SABRE to leverage the PLM's latent knowledge to "probe" for subtle metaphorical cues that conventional span classifiers often overlook. Furthermore, we integrate a **Biaffine Attention** mechanism to strictly enforce the directionality

*Corresponding author.
Our code is available at https://github.com/SeaHai-Li/SABRE.

(Source → Target) of the extracted relations. Finally, to mitigate the severe class imbalance inherent in MRE, a **Focal Loss** objective [7] is incorporated during training to focus on hard examples.

Experimental results on the CMRE dataset demonstrate that SABRE achieves a new state-of-the-art F1 score of **67.04%**, outperforming the baseline by **3.58%**. Notably, SABRE yields a substantial improvement in recall, providing strong empirical evidence that our one-stage, structure-aware design effectively addresses the recall bottleneck of generate-then-classify approaches.

II. Related Work

A. From Metaphor Identification to Relation Extraction

Metaphor processing has long been a central topic in natural language processing. Early studies primarily formulated metaphor identification as a token-level sequence labeling task [8], employing recurrent or convolutional neural networks to detect metaphorical expressions based on local contextual cues [9], [10]. With the emergence of Pre-trained Language Models (PLMs), subsequent work leveraged contextualized representations to capture deeper semantic incongruity, leading to substantial improvements in metaphor detection performance [11]–[14]. Recent research has also expanded to metaphor generation [4], [15], [16].

However, these tasks often lack explicit structural modeling. To address this, Chen et al. [3] formalized *Metaphorical Relation Extraction* (MRE). While their span-based model sets a strong baseline, it relies on a cascaded *generate-then-classify* strategy. This factorization causes error propagation: if implicit metaphorical spans are pruned during the generation phase, they cannot be recovered by the subsequent relation classifier [17]. SABRE aims to resolve this bottleneck by unifying span perception and relation decoding in a one-stage framework.

B. Span-based and Unified Relation Extraction

Relation Extraction (RE) has evolved from modular pipelines to unified, end-to-end models. Span-based approaches, such as SpERT [18], jointly extract entities and relations by classifying span representations. To further handle overlapping relations and improve efficiency, TPLinker [19] introduced a one-stage token-pair linking framework that decodes relations in a single step.

Despite their success, directly applying these models to MRE is non-trivial. Unlike general entity relations, metaphorical relations are often implicit, semantically abstract, and weakly grounded in surface forms. As a result, conventional span classifiers tend to miss subtle metaphor-related spans, exacerbating recall issues. To address this, recent *prompt-based* paradigms, exemplified by GLiNER [5] and GLiREL [6], treat extraction as a schema-following task by injecting label semantics as natural language prompts into the input. SABRE synergizes these paradigms: it adopts a **one-stage architecture** (conceptually similar to TPLinker) to avoid error propagation, while integrating **Semantic Anchors** to explicitly "probe" the PLM for subtle metaphorical cues.

C. Biaffine Attention for Directed Structural Modeling

Deep Biaffine Attention, originally proposed for graph-based dependency parsing [20], has become a standard mechanism for modeling pairwise interactions in structured prediction tasks. By projecting representations into distinct head and dependent spaces, the biaffine transformation naturally captures directional and asymmetric relationships.

Beyond dependency parsing, biaffine mechanisms have been successfully applied to tasks such as nested Named Entity Recognition [21] and Semantic Role Labeling, where structured and directional dependencies are essential. In the context of MRE, the relationship between a *Source* and a *Target* is inherently asymmetric and directional, closely resembling a semantic dependency rather than a symmetric relation. However, existing MRE approaches typically rely on concatenation-based span-pair classifiers, which lack explicit inductive bias for directionality.

To our knowledge, SABRE is the first framework to introduce Biaffine Attention specifically for metaphorical relation extraction. By modeling Source→Target links as directed dependencies, our approach provides a principled solution to the directionality ambiguity that plagues symmetric decoders, thereby enabling more accurate structural inference of metaphorical mappings.

III. Methodology

We propose **SABRE**, a unified one-stage framework designed to capture the structural asymmetry of metaphorical mappings. As illustrated in Fig. 1, SABRE performs span representation learning and relation decoding jointly within a single structured prediction process.

A. Label-Augmented Input Construction

To explicitly query the pre-trained encoder for metaphor-specific semantics, we formulate MRE as a schema-following extraction task. We define a set of **Natural Language Prompts** $\mathcal{L}$ corresponding to the semantic roles in the CMRE schema. Specifically, $\mathcal{L} = \{l_1, \ldots, l_K\}$ consists of descriptive tokens for roles such as "Tenor" and "Vehicle". Given an input sentence $S = \{w_1, \ldots, w_n\}$, we construct the input sequence X as:

$$X = \texttt{[CLS]} \oplus \mathcal{L} \oplus \texttt{[SEP]} \oplus S \oplus \texttt{[SEP]} \tag{1}$$

By prepending these prompts, the global self-attention mechanism of the BERT encoder allows textual tokens in S to interact with the role descriptions in $\mathcal{L}$. These prompts act as *semantic anchors*, aggregating task-specific context to guide the identification of implicit metaphorical spans.

B. Contextual Encoding and Span Representation

The augmented sequence X is fed into a pre-trained encoder (BERT-wwm) to obtain contextualized representations $\mathbf{H} \in \mathbb{R}^{L \times d}$.

To exhaustively recall all potential metaphors, we enumerate all possible spans up to a maximum length W (set to 15). For a candidate span $s_{(i,j)}$ starting at index i and ending at j, we

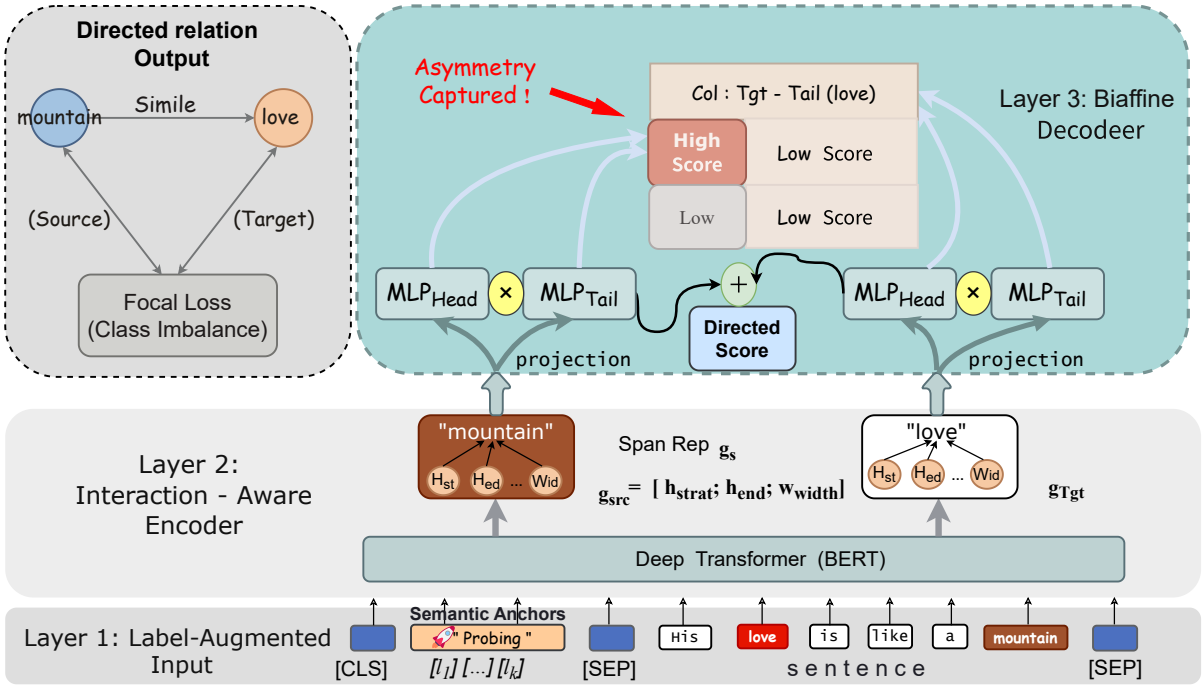

Fig. 1. The architecture of **SABRE**. (1) **Label-Augmented Input**: We prepend semantic anchors to the sentence. (2) **Interaction-Aware Encoder**: The BERT encoder captures dependencies between anchors and text spans via self-attention. (3) **Biaffine Decoder**: A deep biaffine attention mechanism computes directed scores to identify asymmetric metaphorical relations (e.g., Source → Target) effectively.

construct an enhanced span representation $\mathbf{g}_{i,j}$ through the following steps:

1. **Boundary Projection**: To capture boundary-specific semantics, we employ two separate linear projection layers for the start and end tokens:

$$\mathbf{h}'_i = \text{MLP}_{start}(\mathbf{H}_i), \quad \mathbf{h}'_j = \text{MLP}_{end}(\mathbf{H}_j) \tag{2}$$

2. **Width Embedding**: To incorporate structural information, we introduce a learnable width embedding matrix $\mathbf{E}_w$. The width feature is retrieved as $\mathbf{w}_{len} = \mathbf{E}_w[j-i+1]$. 3. **Fusion**: The final span representation is obtained by concatenating these features followed by Layer Normalization and Dropout:

$$\mathbf{g}_{i,j} = \text{LayerNorm}([\mathbf{h}'_i; \mathbf{h}'_j; \mathbf{w}_{len}]) \tag{3}$$

This design ensures that $\mathbf{g}_{i,j}$ encodes both the semantic content and the structural boundary of the span.

C. Biaffine Decoder for Directed Relations

Metaphorical relations are inherently directional (Source → Target). To model this inductive bias, we employ a **Deep Biaffine Attention** mechanism [20]. First, we project the span representation $\mathbf{g}_{i,j}$ into two task-specific subspaces: a "Head" space (representing the Source role) and a "Tail" space (representing the Target role):

$$\mathbf{h}^{(H)} = \text{MLP}_{head}(\mathbf{g}_{i,j}), \quad \mathbf{h}^{(T)} = \text{MLP}_{tail}(\mathbf{g}_{i,j}) \tag{4}$$

Then, the directed relation score between a potential Source span s_i and a Target span s_j is computed as:

$$\text{Score}(s_i, s_j) = (\mathbf{h}_{s_i}^{(H)})^\top \mathbf{U} \mathbf{h}_{s_j}^{(T)} + \mathbf{W}[\mathbf{h}_{s_i}^{(H)}; \mathbf{h}_{s_j}^{(T)}] + \mathbf{b} \tag{5}$$

where $\mathbf{U}$ is the bi-linear parameter matrix. This asymmetric scoring ensures that the model strictly distinguishes between the Source and Target roles.

D. Optimization

Given the extreme sparsity of valid metaphorical pairs, we adopt **Focal Loss** [7] as the training objective to mitigate class imbalance:

$$\mathcal{L} = -\alpha(1-p_t)^\gamma \log(p_t) \tag{6}$$

where p_t is the predicted probability for the ground-truth class. This objective down-weights easy negatives (background pairs) and forces the model to focus on hard metaphorical examples. **The complete training procedure is summarized in Algorithm 1.**

IV. Experiments

A. Experimental Setup

Dataset and Metrics. We conduct experiments on the CMRE dataset [3], utilizing the standard 8:1:1 split. We report Precision (P), Recall (R), and F1-score. Consistent with the baseline [3], we adopt standard evaluation criteria:

- **Span Extraction**: A span is correct if and only if both its boundary and type match the ground truth.
- **Relation Extraction**: A relation is correct if and only if both argument spans are correctly identified and the relation type is correct.

TABLE I
MAIN RESULTS ON CMRE DATASET. THE BASELINE RESULTS ARE BASED ON THE SPERT MODEL [3].

Model	Relation Extraction			Span Extraction		
	P	R	F1	P	R	F1
Baseline (SpERT-based) [3]	**70.71**	57.59	63.46	**78.99**	65.57	71.64
SABRE (Ours)	66.85	**67.22**	**67.04**	77.59	**73.80**	**75.65**
Improvement	*-3.86*	***+9.63***	***+3.58***	*-1.40*	***+8.23***	***+4.01***

Algorithm 1 Training Process of SABRE

Input: Training set $\mathcal{D} = \{(S, R)\}$, Label Prompts $\mathcal{L}$, Max Epochs E, Learning Rate η
Output: Trained Parameters $\theta = \{\theta_{enc}, \theta_{mlp}, \theta_{biaff}, \mathbf{E}_w\}$

Initialize model parameters θ randomly or from pre-trained weights
for $e = 1$ to E **do**
  for each batch (S_{batch}, R_{batch}) in $\mathcal{D}$ **do**
    // 1. Label-Augmented Input Construction (Sec. 3.1)
    Construct $X \leftarrow$ `[CLS]` $\oplus\, \mathcal{L} \oplus$ `[SEP]` $\oplus\, S_{batch} \oplus$ `[SEP]`
    $\mathbf{H} \leftarrow \text{Encoder}(X)$ ▷ Get Contextual Reps
    // 2. Span Representation (Sec. 3.2)
    $\mathcal{G} \leftarrow \emptyset$
    for each candidate span $s_{(i,j)}$ in X **do**
      Retrieve width embedding $\mathbf{w}_{len} \leftarrow \mathbf{E}_w[j - i + 1]$
      Project boundaries: $\mathbf{h}'_i, \mathbf{h}'_j \leftarrow \text{MLP}(\mathbf{H}_i), \text{MLP}(\mathbf{H}_j)$
      $\mathbf{g}_{i,j} \leftarrow \text{LayerNorm}([\mathbf{h}'_i; \mathbf{h}'_j; \mathbf{w}_{len}])$
      Add $\mathbf{g}_{i,j}$ to set $\mathcal{G}$
    end for
    // 3. Biaffine Decoding (Sec. 3.3)
    $\mathbf{h}^{(H)}, \mathbf{h}^{(T)} \leftarrow \text{MLP}_{head}(\mathcal{G}), \text{MLP}_{tail}(\mathcal{G})$
    $Scores \leftarrow \text{Biaffine}(\mathbf{h}^{(H)}, \mathbf{h}^{(T)})$
    // 4. Optimization (Sec. 3.4)
    Calculate loss: $\mathcal{L}_{focal} \leftarrow \text{FocalLoss}(Scores, R_{batch})$
    Update parameters: $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{focal}$
  end for
end for
return θ

Implementation Details. We employ Chinese-BERT-WWM-ext [22] as the encoder backbone. The model is optimized using **AdamW** with a learning rate of **4e-5** and a linear warmup for the first 100 steps. **Configuration.** We set the maximum span width to **15**, strictly adhering to the baseline setting [3] to ensure a fair comparison. We prioritize the official SpERT-based baseline [3] to strictly align with the established benchmark protocols. We exclude token-pair paradigms (e.g., TPLinker [19]) as they require fundamental reformulation of the span-based MRE objective, hindering direct and fair comparison. To address class imbalance, we follow the baseline implementation by employing a negative sampling ratio of **4:1** and optimizing via **Focal Loss** ($\alpha = 0.75, \gamma = 2.0$). All experiments are conducted on a single NVIDIA RTX 4090 GPU.

TABLE II
FINE-GRAINED ANALYSIS. WE COMPARE THE F1 SCORES (%) AGAINST THE BASELINE [3].

Metaphor Type	Baseline (F1)	P	R	SABRE (F1)
Nominal	67.24	74.11	69.75	**71.86**
Verb	55.46	60.00	56.38	**58.14**
Attribute	**36.73**	47.37	25.71	33.33
Simile	68.82	79.17	76.00	**77.55**

B. Main Results

Table I presents the performance comparison between SABRE and the baseline SpERT-based model.

Breaking the Recall Bottleneck. The most significant observation is the dramatic improvement in Recall. While the **generate-then-classify baseline** achieves a high Precision (70.71%), its Recall is limited to 57.59%, confirming the "Fail-to-Recall" issue inherent in cascaded pruning. In contrast, SABRE achieves a **Recall of 67.22%**, an absolute improvement of **+9.63%**. This empirical evidence provides strong support for our hypothesis: the one-stage architecture, combined with Semantic Anchors, effectively recovers implicit metaphors that are prematurely discarded by pipeline approaches.

Overall Performance. Consequently, SABRE achieves a new state-of-the-art F1 score of **67.04%**, outperforming the **strongest baseline** (63.46%) by **3.58%**. Furthermore, in the Span Extraction task, SABRE also demonstrates superior performance (75.65% F1 vs. 71.64% F1), proving that our structure-aware span representation is robust even for boundary detection.

C. Fine-grained Analysis

To understand the source of the significant Recall improvement (+9.63%), we analyze performance across four metaphor categories (Table II).

Breaking the Bottleneck in Structured Metaphors. The global recall gain is primarily driven by *Simile* (**+8.73%** F1) and *Nominal* (**+4.62%** F1) categories.

- For *Similes*, our **Semantic Anchors** effectively capture explicit markers (e.g., "like"), drastically reducing missed detections caused by pruning in pipeline models.

TABLE III
ABLATION STUDY ON RELATION EXTRACTION.

Model Settings	P	R	F1
SABRE (Full)	**66.85**	**67.22**	**67.04**
w/o Semantic Prompt	62.21	65.56	63.84
w/o Biaffine Decoder	63.84	62.78	63.31
w/o Both	62.48	63.52	62.99

- For *Nominals*, which typically follow a Subject-Copula-Object structure, the **Biaffine Decoder** successfully recovers the directed dependency that symmetric classifiers often miss.

The Remaining Challenge: Attribute Metaphors. In contrast to the structural categories, performance on *Attribute* metaphors remains a bottleneck (33.33%), with no significant gain in Recall. This limitation aligns with the baseline findings [3] and stems from two factors:

1) **Data Scarcity**: As noted in the dataset statistics, attribute metaphors are significantly underrepresented in the training set [3]. This extreme sparsity inherently limits the efficacy of loss re-weighting strategies, making it difficult for the model to generalize patterns.
2) **Semantic Abstractness**: Unlike concrete entity spans, attributes (e.g., "heaviness" of mood) often lack clear boundaries. Under strict evaluation metrics, this ambiguity disproportionately affects span-based models compared to token-level tagging.

D. Ablation Study

We conduct an ablation study (Table III) to verify the contribution of each module.

Impact of Semantic Anchors. Removing the semantic prompts (w/o Prompt) causes a **3.20% drop** in Relation F1. Notably, we observe the sharpest decline in Precision (from 66.85% to 62.21%). This indicates that the prompts act as a crucial *semantic filter*: by explicitly querying the encoder with role descriptions, the model can better distinguish true metaphorical expressions from literal analogical phrases, thereby reducing false positives.

Impact of Biaffine Decoder. Replacing the Biaffine decoder with a standard linear classifier (w/o Biaffine) results in the most significant degradation (**-3.73%** F1). Unlike the prompt ablation, this setting suffers primarily from a drop in Recall (from 67.22% to 62.78%). This highlights the necessity of structural modeling. Given that the parameter increase from the Biaffine layer is negligible compared to the encoder backbone, the performance gain is primarily attributed to the inductive bias for modeling directed asymmetry rather than model capacity.

Synergy. Removing both components (w/o Both) reduces the performance to 62.99%, which is comparable to the baseline. This confirms that the combination of prompt-enhanced encoding (for precision) and structure-aware decoding (for recall) is the key driver of our SOTA results.

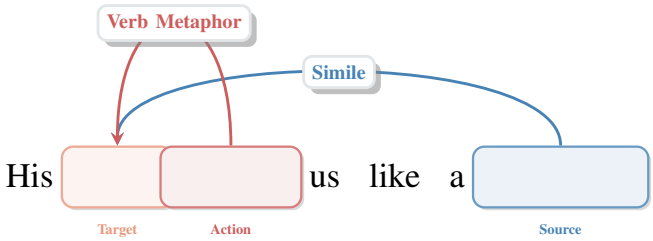

Fig. 2. Case Study: "*His love presses us like a mountain*". This sentence contains a complex multi-relational structure: a **Simile** (Mountain → Love) and a **Verb Metaphor** (Presses → Love). SABRE successfully decodes both overlapping relations, demonstrating the capability of the Biaffine decoder to capture one-to-many dependencies, which is often challenging for pipeline approaches due to error propagation.

E. Qualitative and Error Analysis

To provide a comprehensive understanding of SABRE's capabilities and limitations, we conduct a detailed analysis of specific cases from the test set.

Handling Multi-Metaphor Scenarios: Unlike simple sentences, real-world expressions often contain multiple overlapping metaphors. As visualized in Fig. 2, the sentence "*His love presses us like a mountain*" (ID 8) contains both a Simile (Mountain → Love) and a Verb Metaphor (Presses → Love).

SABRE successfully identifies both relations. This indicates that our **Biaffine Decoder** effectively models the sentence as a directed graph, allowing multiple Source spans to point to the same Target span (One-to-Many mapping). This structural awareness is a significant advantage over **cascaded approaches**, which often struggle with such complex dependencies due to error propagation.

Error Patterns: Despite the performance gains, we analyzed typical errors (Table IV) to identify directions for future improvement:

- *Boundary Over-Extension*: As seen in the first example (ID 196), the model correctly locates the core metaphor but fails to truncate long modifiers (e.g., including "in an

TABLE IV
REPRESENTATIVE ERROR CASES GENERATED BY SABRE.

Error Type	Example Case
Boundary Error	*Input*: "The sky is dark, like [the roof covered with cobwebs in an old house]." (ID 196) *Gold*: "roof covered with cobwebs" *Pred*: "roof covered with cobwebs **in an old house**" (Span too long)
False Positive	*Input*: "With population explosion, how can the country prosper?" (ID 726) *Gold*: explosion → population *Pred*: explosion → population (✓), **prosper → population** (×)
Missed Relation	*Input*: "Human feelings are like flowers blooming and fading." (ID 623) *Gold*: flowers → feelings *Pred*: *None* (Abstract mapping missed)

old house"). This is a common challenge for span-based models, suggesting a need for syntax-aware boundary pruning.

- *Contextual Confusion*: In complex sentences like ID 726, the model occasionally hallucinates relations between abstract concepts (e.g., linking "prosper" to "population"). This implies that while Semantic Anchors are effective, they may sometimes over-trigger in high-density abstract contexts.

V. Conclusion

In this paper, we presented SABRE, a unified end-to-end structured framework that performs joint span representation learning and relation decoding without cascaded pruning for Chinese Metaphorical Relation Extraction. By identifying the limitations of prior generate-then-classify paradigms—specifically their cascaded error propagation and weak directionality modeling—we reformulated MRE as a structure-aware semantic matching task. Our approach integrates **Semantic Anchors** to probe latent linguistic knowledge for implicit span detection and utilizes a **Biaffine Decoder** to rigorously enforce the asymmetric nature of metaphorical mappings.

Empirical results on the CMRE benchmark demonstrate that SABRE not only achieves a new state-of-the-art F1 score of **67.04%** but also delivers a substantial **9.63% improvement in Recall**. This confirms that our end-to-end design effectively breaks the "fail-to-recall" bottleneck inherent in **cascaded extraction frameworks**. In future work, we plan to explore the integration of SABRE with Large Language Models (LLMs) to further enhance the interpretability of extracted metaphorical relations and extend our framework to cross-lingual settings.

Acknowledgment

This research is supported by the Natural Science Foundation of Hebei Province (Grant No. F2026201071).

References

[1] L. Cameron, *Metaphor in educational discourse*. A&C Black, 2003.

[2] G. Lakoff and M. Johnson, *Metaphors we live by*. University of Chicago press, 2008.

[3] G. Chen, T. Wu, M. Cheng, X. Han, J. Gong, S. Wang, and W. Song, "Chinese metaphorical relation extraction: Dataset and models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 9085–9095.

[4] Y. Shao, X. Yao, X. Qu, C. Lin, S. Wang, W. Huang, G. Zhang, and J. Fu, "Cmdag: A chinese metaphor dataset with annotated grounds as cot for boosting metaphor generation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 3357–3366.

[5] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Gliner: Generalist model for named entity recognition using bidirectional transformer," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 5364–5376.

[6] J. Boylan, C. Hokamp, and D. G. Ghalandari, "Glirel-generalist model for zero-shot relation extraction," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 8230–8245.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[8] E.-L. Do Dinh and I. Gurevych, "Token-level metaphor detection using neural networks," in *Proceedings of the Fourth Workshop on Metaphor in NLP*, 2016, pp. 28–33.

[9] M. Rei, L. Bulat, D. Kiela, and E. Shutova, "Grasping the finer point: A supervised similarity network for metaphor detection," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1537–1546.

[10] G. Gao, E. Choi, Y. Choi, and L. Zettlemoyer, "Neural metaphor detection in context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 607–613.

[11] C. Su, F. Fukumoto, X. Huang, J. Li, R. Wang, and Z. Chen, "Deepmet: A reading comprehension paradigm for token-level metaphor detection," in *Proceedings of the second workshop on figurative language processing*, 2020, pp. 30–39.

[12] M. Ge, R. Mao, and E. Cambria, "Explainable metaphor identification inspired by conceptual metaphor theory," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 10 681–10 689.

[13] E. Aghazadeh, M. Fayyaz, and Y. Yaghoobzadeh, "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)*, 2022, pp. 2037–2050.

[14] Y. Li, S. Wang, C. Lin, F. Guerin, and L. Barrault, "Framebert: Conceptual metaphor detection with frame embedding learning," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023, pp. 1558–1563.

[15] Y. Li, C. Lin, and F. Guerin, "Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling," in *Proceedings of the 29th international conference on computational linguistics*, 2022, pp. 6468–6479.

[16] ——, "Nominal metaphor generation with multitask learning," in *Proceedings of the 15th International Conference on Natural Language Generation*, 2022, pp. 225–235.

[17] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2021, pp. 50–61.

[18] M. Eberts and A. Ulges, "Span-based joint entity and relation extraction with transformer pre-training," in *ECAI 2020*. IOS Press, 2020, pp. 2006–2013.

[19] Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun, "Tplinker: Single-stage joint extraction of entities and relations through token pair linking," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1572–1582.

[20] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," in *International Conference on Learning Representations*, 2017.

[21] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6470–6476.

[22] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3504–3514, 2021.