How Sememic Components Can Benefit Link Prediction for Lexico-Semantic Knowledge Graphs?

Hansi Wang[1,2], Yue Wang[1,2], Qiliang Liang[1,2], Yang Liu[1,2]*
[1]Key Laboratory of Computational Linguistics, Ministry of Education, Peking University
[2]School of Computer Science, Peking University
wanghansi2019@pku.edu.cn, wyy209@pku.edu.cn, lql.pkucs@gmail.com,
liuyang@pku.edu.cn

Abstract

Link Prediction (LP) aims to predict missing triple information within a Knowledge Graph (KG). Existing LP methods have sought to improve the performance by integrating structural and textual information. However, for lexico-semantic KGs designed to document fine-grained sense distinctions, these types of information may not be sufficient to support effective LP. From a linguistic perspective, word senses within lexico-semantic relations usually show systematic differences in their sememic components. In light of this, we are motivated to enhance LP with sememe knowledge. We first construct a Sememe Prediction (SP) dataset, SememeDef, for learning such knowledge, and two Chinese datasets, HN7 and CWN5, for LP evaluation; Then, we propose a method, SememeLP, to leverage this knowledge for LP fully. It consistently and significantly improves the LP performance in both English and Chinese, achieving SOTA MRR of 75.1%, 80.5%, and 77.1% on WN18RR, HN7, and CWN5, respectively; Finally, an in-depth analysis is conducted, making clear how sememic components can benefit LP for lexico-semantic KGs, which provides promising progress for the completion of them[1].

1 Introduction

Link Prediction (LP) aims to predict missing (*head*, *relation*, *tail*) triples within a Knowledge Graph (KG) (Dettmers et al., 2018). KGs, whether constructed manually or automatically, often suffer from incomplete knowledge and potential inaccuracies that limit their utility. LP plays a crucial role in addressing these problems across various types of KGs, including lexico-semantic KGs.

Existing LP methods fall into two main categories: embedding-based and Pre-trained Language Model (PLM)-based. Embedding-based methods (Bordes et al., 2013; Dettmers et al., 2018; Balazevic et al., 2019; Chen et al., 2021) optimize concept and relation embeddings using structural information, with various scoring functions to predict missing triples. In contrast, PLM-based methods (Yao et al., 2019; Wang et al., 2022a; Lin et al., 2024; Li et al., 2024) focus on using PLMs to encode available textual descriptions for prediction. However, for lexico-semantic KGs designed to document fine-grained sense distinctions (Choi et al., 2024), structural or textual information may not be sufficient to support effective LP. Thus, it is necessary to explore more fine-grained lexico-semantic knowledge, like sememic components, to further improve the performance.

In lexico-semantic KGs, word senses often serve as the vertices. While treating them as atomic units, current LP methods overlook the effectiveness of internal semantic composition. From a linguistic perspective, word senses can be broken down into smaller units through componential analysis (Lounsbury, 1956; Goodenough, 1956; Lyons, 1968; Leech, 1974), and the minimal indivisible sememic components are called sememes (Bloomfield, 1926). Some linguists propose that word senses in any language can be composed of a finite set of sememes (Dong, 1988; Wierzbicka, 1996). For instance, the primary sense of *boy* can be represented by a combination of sememes: {human, male, immature}. This decomposition potentially provides a systematic framework for representing lexico-semantic relations, for example: Antonymy can be formalized through single-component negation while preserving others (e.g., *boy*: {human, male, immature} → *girl*: {human, female, immature}); Also, hypernymy can be formalized through subset containment (e.g., *boy*: {human, male, immature} → *child*: {human, immature}). In light of this, the sememe information of word senses may offer potential to improve LP for lexico-semantic KGs.

*Corresponding author
[1]The resources and codes for this paper are available at https://github.com/COOLPKU/SememeLP

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14654–14673
November 4-9, 2025

At the same time, practical challenges exist in leveraging sememe information for LP. Nowaday's lexico-semantic KGs seemingly lack deployment of such information, while manual sememe annotation for them would be prohibitively labor-intensive and time-consuming. Moreover, not all word senses in them can acquire sememe information from existing sememe knowledge bases (KBs) like HowNet (Dong et al., 2010). Thus, the introduction of Sememe Prediction (SP) (Xie et al., 2017), aiming to assign suitable sememes for word senses from a pre-defined set, provides a promising approach to tackle these issues, while the relatively low accuracy of SP (Li et al., 2018; Du et al., 2020; Qi et al., 2020, 2022) may largely hinder effective utilization of sememe information.

Considering the potential value of sememic components and practical application issues, in this paper, we are motivated to explore the application of such knowledge to benefit LP. We first construct SememeDef, an SP dataset containing a substantial amount of word sense definitions with sememe annotations; Then, SememeLP, by incorporating embedding-formatted sememe knowledge representations learned from SememeDef into BERT-based models, is designed for LP in KGs; We finally evaluate the method on the English dataset WN18RR (Dettmers et al., 2018), along with HN7 and CWN5, our newly constructed Chinese ones. Experimental results show that SememeLP consistently and significantly improves LP performance across both languages. Further analysis reveals that our sememe knowledge representations generalize well to word senses unseen in SememeDef, facilitating the model to leverage sememe differences between related senses for more accurate predictions, which demonstrates the effectiveness and robustness of the method.

In summary, the main contributions we have achieved are as follows:

(1) We provide the SememeDef dataset for SP, along with two Chinese datasets, HN7 and CWN5, for LP, aiming to alleviate the scarcity of both SP and Chinese LP resources;

(2) We propose the SememeLP method to leverage fine-grained sememe knowledge for enhancing LP in lexico-semantic KGs, achieving SOTA MRR of 75.1%, 80.5%, and 77.1% on WN18RR, HN7 and CWN5, respectively;

(3) We make clear how sememic components can benefit LP for lexico-semantic KGs, providing promising progress for the completion of such KGs, facilitating downstream tasks enriched by them.

(4) We tackle the challenges of leveraging sememe information in annotation-scarce scenarios and present a potentially generalizable method to utilize such information to benefit more lexico-semantic tasks.

2 Related Work

2.1 Link Prediction

Resources: In English, WN18RR (Dettmers et al., 2018), built upon WordNet (Miller et al., 1990), is the widely-used lexico-semantic dataset for LP. In contrast, while some influential lexico-semantic KBs exist in Chinese, such as HowNet (Dong et al., 2010) and Chinese WordNet (CWN) (Huang et al., 2010), there remains an obvious lack of standardized lexico-semantic LP datasets. This limitation impedes the progress of LP for this language.

Methods: Existing LP methods can be broadly categorized into embedding-based and PLM-based methods. Embedding-based methods (Bordes et al., 2013; Trouillon et al., 2016; Dettmers et al., 2018; Sun et al., 2019; Balazevic et al., 2019; Vashishth et al., 2020; Chen et al., 2021; Liu et al., 2022) focus on leveraging structural information in KGs to learn concept/relation representations, while often neglecting textual descriptions. In contrast, PLM-based methods (Yao et al., 2019; Kim et al., 2020; Wang et al., 2021a; Chen et al., 2022) incorporate textual descriptions into PLMs to obtain representations for prediction. Through effective negative sampling strategies (Wang et al., 2022a; Qiao et al., 2023; Lin et al., 2024) and further integration of structural information (Chen et al., 2023; Li et al., 2024), PLM-based methods outperform embedding-based on several LP benchmarks like WN18RR (Dettmers et al., 2018), FB15K-237 (Toutanova and Chen, 2015), and Wikidata5M (Wang et al., 2021b).

From the perspective of information utilization, the above-mentioned methods mainly rely on structural and textual information common across different types of KGs. For lexico-semantic KGs with fine-grained sense distinctions, these types of information may not be sufficient.

2.2 Sememe Prediction

Resources: HowNet (Dong et al., 2010), the most comprehensive sememe KB, provides the foundation for SP research. It comprises 237,974 English and Chinese lexicons annotated with 2,540 expert-

defined sememes (Qi et al., 2019). In HowNet, each word sense is defined by a series of relevant sememes, with the first one as the main sememe representing its core semantic category. For example, the sense *institution dedicated to education* of *school* is annotated with sememes: `institutePlace|场所`, `education|教育`, `study|学习`, and `teach|教`. Whereas, HowNet lacks textual definitions for word senses, which are necessary information for many SP methods (Li et al., 2018; Du et al., 2020; Qi et al., 2022). SememeBabel (Qi et al., 2020), built upon HowNet and a multilingual KB BabelNet (Navigli and Ponzetto, 2010), is another sememe annotation resource specifically built for SP. It contains 15,461 synsets with both textual definitions and sememe annotations, helping to address the issue of definition absence in HowNet for SP training. Nevertheless, while 70,645 English (*definition, sememes*) pairs are extracted, only 8,555 Chinese ones are obtained. The relative scarcity of data may greatly affect the performance of SP in Chinese.

Methods: Existing methods have explored incorporating different types of information into SP models for boosting the performance (Qi et al., 2021), including word sense definitions (Li et al., 2018; Du et al., 2020; Qi et al., 2022), lexico-semantic relations (Qi et al., 2020), multilingual synonyms (Qi et al., 2018, 2022), and even visual information (Qi et al., 2022). Among these, definitions are well-suited for SP as they often align with sememe annotations. For example, the words in the sense *a young male person* of *boy* directly map to its sememes: `immature`, `male`, and `human`. They are also generally more accessible than multilingual and multimodal information in common lexico-semantic KB scenarios.

3 Resources

Considering the scarcity of SP and LP resources for Chinese, we aim to construct a series of datasets to alleviate these issues: We first provide the SememeDef dataset for SP, and then build HN7 and CWN5 for Chinese LP.

3.1 SP Dataset

To address the scarcity of Chinese SP resources[2], we attempt to obtain more (*definition, sememes*) pairs through Word Sense Alignment (WSA) between HowNet and the Contemporary Chinese Dictionary (CCD)[3]. Due to different sense granularities across resources (Matuschek and Gurevych, 2014), rigorous WSA needs to consider various mapping scenarios (e.g., one-to-one, one-to-many, many-to-one, and many-to-many). This complexity poses challenges for existing automated methods (Ji et al., 1998; Matuschek and Gurevych, 2013; Pilehvar and Navigli, 2014; Yao et al., 2021).

Considering the practical demand for data expansion of SP, we focus on identifying semantically consistent (i.e., one-to-one) word sense pairs across HowNet and CCD. To assess this consistency, inspired by recent advances in Entity Alignment (EA) (Jiang et al., 2024; Chen et al., 2024a,b), we employ three Large Language Models (LLMs): Qwen2.5-72B-Instruct (Qwen et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024), and Yi-Lightning (Wake et al., 2024)[4]. These LLMs score each sense pair on a 5-point scale (1: low consistency to 5: high consistency) based on the information from both resources. Detailed configurations are provided in Appendix A.

After manually checking the LLM outputs, we select sense pairs scored at least 4 by all the LLMs to expand the Chinese SP data. And we also conduct human evaluation (detailed in Appendix B) to further validate the reliability of these LLM-scored pairs. Results show that there is substantial annotation consistency between LLMs and human annotators, as evidenced by comparable pairwise agreements (i.e., human-human: 0.935–0.960; human-LLM: 0.930–0.975).

The resulting dataset, named SememeDef, contains 70,645 English samples and 43,163 Chinese ones, covering 2,042 and 1,762 sememes, respectively. Table 1 shows the examples from each language.

3.2 Chinese Lexico-Semantic LP Datasets

To facilitate LP evaluation in Chinese, two datasets, HN7 and CWN5, are built upon HowNet and CWN, respectively. Considering the existing research progress (Wang et al., 2025) in sense definition similarity computation, it is relatively less challenging to mine the synonymy relation compared with other relation types. Therefore, following

[2]We do not adopt LLM-generated sememe annotations, given that it is significantly challenging for LLMs to identify appropriate sememes from over 2,000 candidates.

[3]The most authoritative and influential Chinese dictionary, published by the Commercial Press.

[4]The selection of these models balances their performance in Chinese understanding with API costs.

Word Sense Definition	Sememes	Main Sememe
a person whose job is teaching	humanl人, occupationl职位, educationl教育, teachl教	humanl人
专门进行教育的机构 (*institution dedicated to education*)	institutePlacel场所, educationl教育, studyl学习, institutePlacel场所 teachl教	

Table 1: Examples from SememeDef, our newly constructed SP dataset, where English and Chinese word sense definitions are shown with their corresponding sememe information.

Dataset	Head Synset	Relation	Tail Synset
HN7	{幼, 少 (*young*)}	antonymy	{年迈, 老 (*old*)}
CWN5	{妇女, 女人, (*woman*)}	antonymy	{丁, 男人 (*man*)}

Table 2: Examples from HN7 and CWN5, our newly constructed LP datasets for Chinese.

WN18RR, synsets are set as the vertices in KGs.

For HN7, we construct synsets and extract relations following the instructions of OpenHowNet (Qi et al., 2019). The resulting dataset contains 10,939 synsets and 25,672 triples across 7 relation types. For each synset, GPT-4o (OpenAI et al., 2024) is utilized to generate a unified definition based on sememe information and associated definitions from CCD (detailed in Appendix C). The generated definitions are manually checked to ensure quality.

For CWN5, we directly extract the synsets and relations in CWN. It contains 3,149 synsets and 5,395 triples across 5 relation types.

Table 2 shows examples of antonymy from HN7 and CWN5. More examples and details about these datasets are provided in Appendix D.

4 Methodology

4.1 Task Formulation

A lexico-semantic KG is a directed graph, where the vertices $\mathcal{V}$ are word senses (typically grouped into synsets) linked by various lexico-semantic relations. Each edge in the KG can be denoted by a triple (h, r, t), where h, r, and t represent the head vertice, relation, and tail vertice, respectively.

In this paper, LP aims to predict missing triples in a lexico-semantic KG. It is made up of two subtasks: tail and head prediction. Under the widely adopted evaluation protocol (Wang et al., 2022a), tail prediction $(h, r, ?)$ requires ranking all vertices given h and r, similarly for head prediction$(?, r, t)$. Following previous research (Wang et al., 2022a; Lin et al., 2024; Li et al., 2024), for each (h, r, t), we need to add (t, r^{-1}, h), where r^{-1} denotes the inverse relation of r. This allows unified handling of both types of prediction through tail ranking.

4.2 SememeLP: Enhance LP by Leveraging Sememe Knowledge

We propose SememeLP, a novel method that leverages sememe knowledge to enhance LP for lexico-semantic KGs. SememeLP utilizes a three-stage fusion module to combine sememe features with other features for more robust knowledge representations. In this subsection, we first introduce the overall architecture of SememeLP, then detail the acquisition of sememe knowledge needed by it, and finally present optimization strategies for further improving the training effectiveness and prediction performance.

4.2.1 Overall Architecture

The overall architecture of SememeLP is shown in Figure 1. Specifically, for a candidate triple (h, r, t), we incorporate textual descriptions (i.e., word sense definitions) of (h, r) and t into separate BERT-based encoders, $\boldsymbol{E}_{hr}$ and $\boldsymbol{E}_t$, respectively. The vanilla representations, $\boldsymbol{r}_v^{hr}$ and $\boldsymbol{r}_v^t \in \mathbb{R}^l$, are obtained by pooling the last hidden states.

To derive sememe features of vertices without sememe annotations, a BERT encoder $\boldsymbol{E}_s$, finetuned on SememeDef, is utilized to obtain two types of sememe knowledge representations from their definitions: all-sememe representation $\boldsymbol{r}_a \in \mathbb{R}^l$ encoding all sememes and main-sememe representation $\boldsymbol{r}_m \in \mathbb{R}^l$ encoding the main sememe.

Subsequently, these two types of sememe representations are fused with the vanilla representation. However, there may exist potential challenges in integrating them: First, the main-sememe and all-sememe representations might contain noise due to potential inaccuracies of SP, which degrades the effectiveness of the final representations; Second, the contribution of main-sememe and all-sememe representations for LP may dynamically change due to their own effectiveness or different relation types, while the assignment of static weights usu-

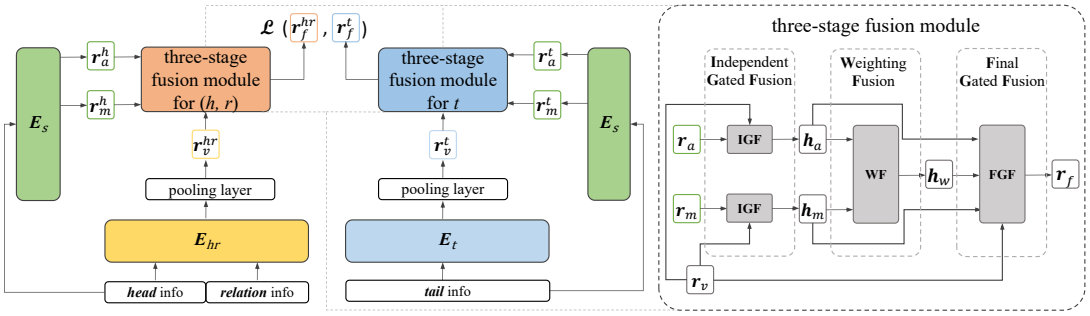

Figure 1: Illustration of the proposed SememeLP. By employing a three-stage fusion module to combine sememe features with other features, SememeLP can generate more robust representations to enhance LP.

ally fails to deal with such variations; Third, sememe features are complementary to others, and their relative importance varies across different LP scenarios. Therefore, it is necessary to assess their importance for more robust final representations dynamically.

To address these challenges, we design a **three-stage fusion module**:

Independent Gated Fusion (IGF) judges the effectiveness of main-sememe and all-sememe representations, and uses vanilla representations to refine them. Formally, two gates with the same architecture are used to fuse $\boldsymbol{r}_v$ with each sememe representation:

$$\boldsymbol{h}_i = \boldsymbol{g}_i \odot \boldsymbol{r}_v + (\boldsymbol{1} - \boldsymbol{g}_i) \odot \boldsymbol{r}_s, \tag{1}$$

where $\boldsymbol{g}_i \in \mathbb{R}^l = \sigma(\text{MLP}([\boldsymbol{r}_v; \boldsymbol{r}_s]))$, and $\boldsymbol{r}_s \in \{\boldsymbol{r}_a, \boldsymbol{r}_m\}$, with $[;]$ denoting concatenation, and σ denoting a sigmoid function. The outputs for $\boldsymbol{r}_a$ and $\boldsymbol{r}_m$ are denoted as $\boldsymbol{h}_a$ and $\boldsymbol{h}_m$, respectively.

Weighting Fusion (WF) performs a weighted fusion of main-sememe and all-sememe representations for the final sememe knowledge representation. Formally, a weighting layer is employed for combining $\boldsymbol{h}_a$ and $\boldsymbol{h}_m$:

$$\boldsymbol{h}_w = w_a \boldsymbol{h}_a + w_m \boldsymbol{h}_m, \tag{2}$$

where $[w_a, w_m] = \text{softmax}(\text{MLP}([\boldsymbol{h}_a; \boldsymbol{h}_m]))$.

Final Gated Fusion (FGF) combines sememe features with other features for the final head-relation and tail representations. Formally, a gate is used for the final fusion of $\boldsymbol{r}_v$ with $\boldsymbol{h}_w$:

$$\boldsymbol{r}_f = \boldsymbol{g}_f \odot \boldsymbol{r}_v + (\boldsymbol{1} - \boldsymbol{g}_f) \odot \boldsymbol{h}_w, \tag{3}$$

where $\boldsymbol{g}_f \in \mathbb{R}^l = \sigma(\text{MLP}([\boldsymbol{r}_v; \boldsymbol{h}_w; \boldsymbol{h}_a; \boldsymbol{h}_m]))$.

Through different fusion modules, we combine $\boldsymbol{r}_v^{hr}$ with the sememe knowledge representations of h, and $\boldsymbol{r}_v^t$ with those of t. Despite sharing the same architecture, these modules are designed for different functions: the (h, r) module learns more accurate mapping to the representation of gold tail, while the t module enriches the representation of t.

Subsequently, the prediction score of (h, r, t) is computed as the cosine similarity of $\boldsymbol{r}_f^{hr}$ and $\boldsymbol{r}_f^t$:

$$f(h, r, t) = \frac{\boldsymbol{r}_f^{hr} \cdot \boldsymbol{r}_f^t}{||\boldsymbol{r}_f^{hr}|| \cdot ||\boldsymbol{r}_f^t||} \tag{4}$$

4.2.2 Encoder for Sememe Knowledge Representation

To obtain sememe knowledge representations, we leverage definitions in the task of SP, as they often align with sememe annotations for word senses. A BERT-based model is used as the sememe knowledge encoder $\boldsymbol{E}_s$, and fine-tuned on two SP tasks: All-Sememe Prediction (ASP), for predicting all sememes of a word sense, and Main-Sememe Prediction (MSP), for predicting its main sememe.

Given a word sense definition d, we design an input template with soft prompts (Hambardzumyan et al., 2021; Qin and Eisner, 2021; Wang et al., 2022b): "`[CLS] [A1] [A2] ... [AL] [ASP] [M1] [M2] ... [ML] [MSP]` d `[SEP]`", where `[A1]`-`[ML]` are learnable template tokens, with `[ASP]` and `[MSP]` as classification tokens for ASP and MSP, respectively. The last hidden states of `[ASP]` and `[MSP]`, denoted by $\boldsymbol{h}_{\texttt{[ASP]}}$ and $\boldsymbol{h}_{\texttt{[MSP]}}$, are used as the all-sememe representation $\boldsymbol{r}_a$ and main-sememe representation $\boldsymbol{r}_m$.

For ASP, we employ a multi-label classifier:

$$\boldsymbol{p}_{asp} = \sigma(\mathbf{W}_{asp} \boldsymbol{h}_{\texttt{[ASP]}} + \boldsymbol{b}_{asp}), \tag{5}$$

where $\mathbf{W}_{asp}$ is a weight matrix, and $\boldsymbol{b}_{asp}$ is a bias vector. The obtained $\boldsymbol{p}_{asp} \in \mathbb{R}^{|\mathcal{S}|}$ contains prediction scores for sememes in a pre-defined sememe

set $\mathcal{S}$. The loss function for ASP is:

$$\mathcal{L}_{asp} = -\frac{1}{|\mathcal{S}|}\left[\sum_{s\in\mathcal{S}_d}\log p_s + \sum_{s\notin\mathcal{S}_d}\log(1-p_s)\right], \tag{6}$$

where p_s is the prediction score of s from $\boldsymbol{p}_{asp}$, and $\mathcal{S}_d$ denotes the ground-truth sememe set.

For MSP, a multi-class classifier is used:

$$\boldsymbol{p}_{msp} = \text{softmax}(\mathbf{W}_{msp}\boldsymbol{h}_{\texttt{[MSP]}} + \boldsymbol{b}_{msp}), \tag{7}$$

where $\mathbf{W}_{msp}$ is a weight matrix, and $\boldsymbol{b}_{msp}$ is a bias vector. The obtained $\boldsymbol{p}_{msp} \in \mathbb{R}^{|\mathcal{S}|}$ contains prediction scores for all sememes in $\mathcal{S}$. The loss function for MSP is:

$$\mathcal{L}_{msp} = -\log p_{ms}, \tag{8}$$

where p_{ms} is the prediction score for the ground-truth main sememe ms.

The final loss function for SP is formulated as:

$$\mathcal{L}_{sp} = \alpha\mathcal{L}_{asp} + (1-\alpha)\mathcal{L}_{msp}, \tag{9}$$

where $\alpha \in [0, 1]$ controls the task weighting.

For analysis convenience, we also describe how to obtain sememe labels predicted by $\boldsymbol{E}_s$ as detailed in Appendix E, while these labels are not directly used by SememeLP.

4.2.3 Optimization by Previous LP Methods

Previous studies have demonstrated that effective contrastive learning and further integration of structural information are crucial for the capability improvement of PLM-based LP methods. Enlightened by this insight, we enhance SememeLP with strategies from two top-performing models: SimKGC (Wang et al., 2022a) and MoCoKGC (Li et al., 2024). From SimKGC, we adopt: (1) three negative sampling strategies (in-batch, pre-batch, and self-negatives) for effective contrastive learning; (2) graph-based re-ranking to leverage structural information. From MoCoKGC, we adopt: (1) momentum tail encoder and tail queue for negative sampling; (2) neighborhood prompts to incorporate structural information; (3) relation prompts to enhance the inferential capabilities of $\boldsymbol{E}_{hr}$. The two variants are named SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$.

For LP training, we use InfoNCE (Oord et al., 2018) loss with additive margin (Yang et al., 2019):

$$\mathcal{L}_{lp} = -\log\frac{e^{(f(h,r,t^*)-\gamma)/\tau}}{e^{(f(h,r,t^*)-\gamma)/\tau} + \sum_{i=1}^{|\mathcal{N}|} e^{f(h,r,t_i')/\tau}}, \tag{10}$$

Method	WN18RR			
	MRR	Hits@1	Hits@3	Hits@10
TransE$^{\diamond}$	24.3	4.3	44.1	53.2
ConvE†	45.6	41.9	47.0	53.1
RotatE$^{\diamond}$	47.6	42.8	49.2	57.1
CompGCN	47.9	44.3	49.4	54.6
HittER	50.3	46.2	51.6	58.4
KG-BERT	21.6	4.1	30.2	52.4
StAR	40.1	24.3	49.1	70.9
CSProm-KG	57.5	52.2	59.6	67.8
SimKGC	67.1	58.5	73.1	81.7
StructKGC	69.6	62.3	74.1	82.7
MoCoKGC	74.2	66.5	79.2	88.1
SememeLP$_{\text{Sim}}$	68.2 (+1.6)	60.3 (+3.1)	73.3 (+0.3)	82.1 (+0.5)
SememeLP$_{\text{MoCo}}$	**75.1** (+1.2)	**67.6** (+1.7)	**79.8** (+0.8)	**88.5** (+0.5)

Table 3: Main results (%) on WN18RR, where † denotes the results from Wang et al. (2021a), and $^{\diamond}$ from Chen et al. (2023). The overall best results are shown in **bold**, with the best results in each category underscored. For SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$, the percentage improvements over SimKGC and MoCoKGC, respectively, are shown in parentheses, with darker colors indicating larger performance gains.

where $\gamma > 0$ is the margin coefficient that encourages higher scores for the correct triple (h, r, t^*), $\tau \in [0, 1]$ is a learnable temperature parameter, and $\mathcal{N}$ is the negative sample set.

5 Experiments

5.1 Experimental Settings

Datasets: In the SP task, we separately split Chinese and English SememeDef data into training and validation sets by 19:1. In LP, three benchmark datasets, WN18RR, HN7, and CWN5, are utilized. For WN18RR, textual information is obtained from Yao et al. (2019). For HN7 and CWN5, following WN18RR, inverse relation test leakage (Dettmers et al., 2018) is prevented by selecting one relation type from each inverse pair (e.g., either hypernymy or hyponymy). Despite removing some relation types, the addition of r^{-1} for head prediction ensures that the resulting data still covers all of them. The data is then split into training, validation, and test sets by 8:1:1. Detailed statistics are shown in Appendix F.

Baselines: On WN18RR, SememeLP is compared against two types of methods, including embedding-based TransE (Bordes et al., 2013), ConvE (Dettmers et al., 2018), RotatE (Sun et al., 2019), CompGCN (Vashishth et al., 2020), HittER (Chen et al., 2021), and PLM-based KG-BERT (Yao et al., 2019), StAR (Wang et al., 2021a),

Method	HN7				CWN5			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
CompGCN	40.4	35.2	42.0	45.4	40.2	33.9	42.7	49.8
HittER	48.3	41.5	50.2	58.0	42.0	34.7	46.5	50.0
CSProm-KG	68.6	66.5	70.5	72.2	49.7	47.5	50.9	53.1
SimKGC	77.4	72.4	80.2	86.9	73.3	66.8	76.9	85.7
StructKGC	77.3	71.4	80.9	88.0	75.6	69.2	80.0	87.7
MoCoKGC	76.5	70.0	80.6	88.7	75.7	68.3	81.7	89.4
SememeLP$_{\text{Sim}}$	78.8 (+1.8)	74.3 (+2.6)	80.6 (+0.5)	87.1 (+0.2)	75.0 (+2.3)	69.1 (+3.4)	78.5 (+2.1)	86.3 (+0.7)
SememeLP$_{\text{MoCo}}$	**80.5** (+5.2)	**74.6** (+6.6)	**84.0** (+4.2)	**91.8** (+3.5)	**77.1** (+1.8)	**69.2** (+1.3)	**82.5** (+1.0)	**90.6** (+1.3)

Table 4: Main results (%) on HN7 and CWN5. The notations follow Table 3.

CSProm-KG (Chen et al., 2023), SimKGC (Wang et al., 2022a), StructKGC (Lin et al., 2024), MoCoKGC (Li et al., 2024). On HN7 and CWN5, we select top-performing methods on WN18RR as baselines.

Evaluation Metrics: Following previous research, we adopt the commonly used evaluation metrics, including MRR, Hits@1, Hits@3, and Hits@10. They are reported under the filtered setting (Bordes et al., 2013), and computed by averaging over two directions, head prediction and tail prediction.

Experimental Configuration: We fine-tune sememe knowledge encoders on SememeDef, selecting the best checkpoints to enhance LP based on validation results[5]. During LP fine-tuning, to reasonably evaluate the contribution of sememe knowledge, the optimal parameter setup for SimKGC and MoCoKGC is first searched for and determined. Then, SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$ adopt the same settings of overlapping hyperparameters as SimKGC and MoCoKGC, respectively. For further details, please refer to Appendix G.

5.2 Main Results

The main test results are shown in Table 3 and 4. From them, we have the following observations:

(1) From the overall results, SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$ achieve consistent improvements on all datasets over SimKGC and MoCKGC, respectively. Notably, SememeLP$_{\text{MoCo}}$ achieves the best performance across all evaluation metrics, with significant improvements of 0.9, 3.1, and 1.4 MRR points on WN18RR, HN7, and CWN5, respectively. This significant performance largely benefits from sememe knowledge, which provides a systematic framework for representing lexico-semantic relations and helps learn more accurate head-to-tail mappings across different relations;

(2) Among different evaluation metrics, our method shows larger improvements in Hits@1 compared to Hits@3 and Hits@10 (e.g., 1.7 vs. 0.8 and 0.5 on WN18RR for SememeLP$_{\text{MoCo}}$). This indicates that the incorporation of more fine-grained lexico-semantic knowledge particularly enhances the model's ability to make more precise predictions among semantically similar candidates;

(3) Among different datasets, our method achieves larger improvements on HN7 compared to the others (e.g., 5.2 vs. 1.2 and 1.8 on MRR for SememeLP$_{\text{MoCo}}$). This is because relations between word senses in HowNet are extracted based on their sememe information. This intrinsic relationship enables the model to leverage sememe differences between related word senses more effectively for performance improvement.

6 Analysis

The significant performance of SememeLP demonstrates that sememe knowledge largely benefits LP for lexico-semantic KGs. To further investigate the effectiveness and underlying mechanisms of the method, detailed analyses are conducted to address the following questions: (1) How does SememeLP perform on LP across different relation types? (2) How robust are the sememe knowledge representations? (3) Why does the incorporation of sememe knowledge enhance LP? (4) What is the contribution of different components in SememeLP?

6.1 Analysis on the Effectiveness Across Different Relation Types

To better understand the overall performance, we conduct a fine-grained analysis on the results across different relation types. As shown in Table 5, SememeLP shows improvement trends across all re-

[5]The best checkpoint for English achieves 68.2% MAP on ASP and 64.4% F1 on MSP, while the Chinese one achieves 74.1% and 67.9%.

Dataset	Method	Anto.	Hype.	Hypo.	Holo.	Mero.	Othr.
WN18RR	MoCoKGC	-	64.1	59.7	**64.6**	60.2	90.5
	SememeLP	-	**65.6**	**60.6**	64.0	**60.4**	**91.4**
HN7	MoCoKGC	92.5	82.0	39.7	**78.7**	40.0	55.8
	SememeLP	**95.3**	**86.2**	**55.1**	74.3	**45.8**	**57.9**
CWN5	MoCoKGC	85.5	**78.0**	**67.7**	46.7	32.0	-
	SememeLP	**91.9**	76.7	64.8	**57.3**	**34.7**	-

Table 5: MRR results (%) across subsets for different relation types. Anto., Hype., Hypo., Holo., Mero., and Othr. are abbreviations of antonymy, hypernymy, hyponymy, holonymy, meronymy, and others, respectively. SememeLP denotes SememeLP$_{\text{MoCo}}$.

Dataset	Method	TT	TF	FT	FF
WN18RR	MoCoKGC	71.2	71.3	76.2	75.2
	SememeLP	72.3 (+1.5)	71.6 (+0.4)	77.1 (+1.2)	75.9 (+0.9)
HN7	MoCoKGC	78.3	78.4	73.4	77.6
	SememeLP	83.4 (+6.5)	82.9 (+5.7)	76.6 (+4.4)	80.5 (+3.7)
CWN5	MoCoKGC	68.4	67.2	65.8	82.5
	SememeLP	71.0 (+3.8)	67.8 (+0.9)	66.9 (+1.7)	83.5 (+1.2)

Table 6: MRR results (%) across different definition similarity scenarios. In column headers, the first T/F denotes whether the head definition is highly similar to some definition in SememeDef, and the second for the tail. SememeLP denotes SememeLP$_{\text{MoCo}}$.

lation types[6], while slight decreases are observed in specific subsets on certain datasets. We further discuss these cases by error analysis in Appendix I.

Notably, hypernym and holonym prediction outperform hyponym and meronym prediction across all datasets. This is due to hypernyms and holonyms usually having unique answers, while hyponyms and meronyms often involve many plausible answers that would confuse the model.

6.2 Analysis on the Robustness of Sememe Knowledge Representations

Considering that synset definitions in KGs may appear in the training set of SememeDef, we further analyze the robustness of sememe knowledge representations across different definition similarity scenarios. Specifically, test triples are divided into four groups according to whether their head/tail definitions are highly similar (Jaccard Similarity > 0.5) to some definitions from SememeDef.

As shown in Table 6, SememeLP$_{\text{MoCo}}$ achieves larger performance gains over MoCoKGC in TT, attributed to more accurate sememe knowledge representations due to higher similarity with the training data. While the gains in other groups are relatively smaller, SememeLP$_{\text{MoCo}}$ still outperforms MoCoKGC. The consistent improvements validate the robustness of our method when confronted with the lack of sememe annotations in KGs.

[6]Relations in WN18RR are categorized for clarity. Detailed information is shown in Appendix H.

Dataset	Group	Anto.	Hype.	Hypo.
WN18RR	Correct	-	42.1	37.4
	Error	-	30.7	33.9
HN7	Correct	27.1	64.6	78.5
	Error	19.5	40.0	65.2
CWN5	Correct	20.5	63.6	66.8
	Error	16.6	41.2	42.2

Table 7: Proportions (%) of samples satisfying sememe difference patterns summarized by us. Correct/Error denotes the group of samples where SememeLP$_{\text{MoCo}}$ ranks the ground-truth tail first or not.

6.3 Analysis on Why Sememe Knowledge Enhances LP

Word senses within lexico-semantic relations usually show systematic differences in the sememe composition. Given the finite nature of sememes, these compositional differences can be captured, learned, and generalized feasibly. Intuitively, incorporating sememe knowledge can help models learn more accurate head-to-tail mappings.

To verify this hypothesis, we analyze test samples with different top-1 predictions by MoCoKGC and SememeLP$_{\text{MoCo}}$, and divide them into two groups based on the correctness. We then examine whether the predicted sememes for the head and gold tail conform to the following difference patterns: (1) For word senses within antonymy relation, there exists only one pair of sememes within antonymy relation in their sememe annotations, while the other sememes are identical; (2) For word senses within hypernymy/hyponymy relation, the sememes of hypernym are contained within the sememes of hyponym. As shown in Table 7, the Correct group consistently exhibits higher percentages of samples satisfying these patterns than the Error group. This indicates that the performance gains of our method primarily stem from samples satisfying systematic sememe differences, indirectly demonstrating that by introducing such knowledge, the model can learn to leverage these differences to make reasonable predictions. To further illustrate the contribution of sememe knowledge, we also conduct case studies in Appendix I.

Method	WN18RR	HN7	CWN5
SememeLP$_{\text{MoCo}}$ w/o *all*	74.4	79.1	75.8
SememeLP$_{\text{MoCo}}$ w/o *main*	74.7	80.3	**77.1**
SememeLP$_{\text{MoCo}}$	**75.1**	**80.5**	**77.1**

Table 8: MRR (%) results of ablation studies for sememe knowledge representations.

Method	ALL	TT	TF	FT	FF
BERT	74.6	**72.5**	71.0	76.8	75.3
SememeLP w/ MLP fusion	74.1	71.8	70.2	76.5	75.0
SememeLP w/o IGF	74.3	71.8	71.2	76.5	74.9
SememeLP w/o WF	74.9	72.4	**71.8**	77.0	75.7
SememeLP w/o FGF	74.9	72.0	71.5	76.8	**75.9**
SememeLP	**75.1**	72.3	71.6	**77.1**	**75.9**

Table 9: MRR (%) results on WN18RR for various fusion methods across different definition similarity scenarios (similar to Section 6.2). In column headers, the first T/F denotes whether the head definition is highly similar to some definition in SememeDef, and the second for the tail. SememeLP denotes SememeLP$_{\text{MoCo}}$.

6.4 Ablation Studies

Sememe Knowledge Representations: To investigate the effectiveness of all-sememe and main-sememe representations, we conduct ablation studies by removing each individually. As shown in Table 8, both of them enhance the LP performance, with the all-sememe representation contributing to more significant improvements across all datasets. This can be attributed to its ability to capture comprehensive features that help discriminate among semantically similar candidates. In contrast, the main-sememe representation can only provide beneficial category constraints on predictions.

Fusion Module: To evaluate the necessity of the three-stage fusion module, we compare it against the following baselines: (1) a BERT-based one that jointly feeds sememe tokens and definitions into BERT; (2) replacing the three-stage fusion module with an MLP-based one. Additionally, we attempt to simplify the three-stage fusion by removing each submodule, respectively, with the specific modifications as follows: a) removing IGF; b) removing WF and adopting equal weights; c) removing FGF and directly summing up the sememe and vanilla representation.

As shown in Table 9, all these simpler methods show inferior overall performance compared to the three-stage fusion, demonstrating that our fusion strategy effectively combines complementary features for performance improvement. Notably, three baselines lacking IGF show significant performance drops in the FF subset, where sememe features tend to be less accurate than in other subsets. This is because SememeLP occasionally fails to filter out the noise in features without IGF, thereby degrading the effectiveness of final representations. This further demonstrates that our fusion strategy helps enhance the model's robustness.

7 Conclusions

This paper is dedicated to revealing how sememic components can benefit LP for lexico-semantic KGs. We first construct an SP dataset, SememeDef, for learning such knowledge, and two Chinese benchmarks, HN7 and CWN5, for LP evaluation. Then, we propose a method, SememeLP, to fully leverage this knowledge for LP. It consistently and significantly improves the performance across both English and Chinese, achieving SOTA MRR of 75.1%, 80.5%, and 77.1% on WN18RR, HN7, and CWN5, respectively. Finally, an in-depth analysis is conducted, revealing that SememeLP can leverage systematic sememic component differences between related word senses to improve prediction accuracy. Our work provides promising progress for the completion of lexico-semantic KGs, facilitating downstream tasks enriched by them.

In the near future, we will explore more approaches to utilize sememe knowledge for enhancing LP and investigate how this kind of knowledge can benefit more lexico-semantic tasks, such as Lexical Relation Classification (Ushio et al., 2021; Pitarch et al., 2023), Lexical Entailment (Moskvoretskii et al., 2024a,b), and Word Sense Disambiguation (Hou et al., 2020; Wang et al., 2024), particularly in annotation-scarce scenarios.

Limitations

Despite achieving significant results on LP, there remain some limitations of our method as follows:

(1) The performance of SememeLP is limited by the effectiveness of sememe knowledge representations. This is demonstrated by the evaluation results for LP, where SememeLP shows more significant improvements on Chinese datasets compared to English, attributed to the better performance of Chinese SP than English. In low-resource languages, the sememe knowledge encoder may be less accurate for modeling such knowledge, thereby weakening the effectiveness of SememeLP;

(2) The inconsistencies of taxonomy between the lexico-semantic KG and sememe KB may also affect the performance. This is demonstrated by the results on the hypernymy and hyponymy subsets of CWN5, where SememeLP makes reasonable but incorrect predictions due to the influence of the HowNet taxonomy (detailed in Appendix I). Consequently, in other benchmarks whose taxonomies differ somewhat from HowNet, while SememeLP makes reasonable predictions, its advantages may not be clearly reflected in the evaluation metrics. More appropriate evaluation metrics remain to be explored and developed;

(3) Regarding the efficiency of training, SememeLP requires fine-tuning an additional PLM to represent sememe knowledge and relies on a fusion module to integrate such representation with other representations. This leads to more training time and memory consumption compared to baseline methods (detailed in Appendix J).

(4) SememeLP focuses on applying sememe knowledge to enhance LP for lexico-semantic KGs, while the effectiveness in factual KGs (e.g., biomedical KGs) remains underexplored. However, there are some gaps between the two types of KGs, which may weaken SememeLP's applicability: In lexico-semantic KGs, nodes (word senses) can be directly decomposed into sememes. In contrast, nodes in factual KGs often represent senses of compound words or phrases (e.g., *gene mutation*), which may require a multi-step decomposition (first into word senses, and then sememes). Considering this gap, it is necessary to incorporate a multi-step semantic decomposition into SememeLP to further boost its applicability in such domains.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No. 62036001).

References

Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China. Association for Computational Linguistics.

Leonard Bloomfield. 1926. A set of postulates for the science of language. *International Journal of American Linguistics*, 15:195 – 202.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems*.

Chen Chen, Yufei Wang, Bing Li, and Kwok-Yan Lam. 2022. Knowledge is flat: A Seq2Seq generative framework for various knowledge graph completion. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4005–4017, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chen Chen, Yufei Wang, Aixin Sun, Bing Li, and Kwok-Yan Lam. 2023. Dipping PLMs sauce: Bridging structure and text for effective knowledge graph completion via conditional soft prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11489–11503, Toronto, Canada. Association for Computational Linguistics.

Sanxing Chen, Xiaodong Liu, Jianfeng Gao, Jian Jiao, Ruofei Zhang, and Yangfeng Ji. 2021. HittER: Hierarchical transformers for knowledge graph embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10395–10407, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. 2024a. Entity alignment with noisy annotations from large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 15097–15120. Curran Associates, Inc.

Xuan Chen, Tong Lu, and Zhichun Wang. 2024b. Llm-align: Utilizing large language models for entity alignment in knowledge graphs. *CoRR*, abs/2412.04690.

Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karen Fort, and Bruno Guillaume. 2024. Beyond model performance: Can link prediction enrich French lexical graphs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2329–2341, Torino, Italia. ELRA and ICCL.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 81 others. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zhendong Dong. 1988. Knowledge description: what, how and who. In *Proceedings of International Symposium on Electronic Dictionary*, volume 18.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. HowNet and its computation of meaning. In *Coling 2010: Demonstrations*, pages 53–56, Beijing, China. Coling 2010 Organizing Committee.

Jiaju Du, Fanchao Qi, Maosong Sun, and Zhiyuan Liu. 2020. Lexical sememe prediction using dictionary definitions by capturing local semantic correspondence. *CoRR*, abs/2001.05954.

Ward H Goodenough. 1956. Componential analysis and the study of meaning. *Language*, 32(1):195–216.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised Chinese word sense disambiguation method based on HowNet. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1752–1757, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juren Huang, Shukai Xie, Jiafei Hong, Yunzhu Chen, Yili Su, Yongxiang Chen, and Shengwei Huang. 2010. Chinese wordnet: design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese information processing*, 24(02):14–23.

Donghong Ji, Junping Gong, and Changning Huang. 1998. Combining a Chinese thesaurus with a Chinese dictionary. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 600–606, Montreal, Quebec, Canada. Association for Computational Linguistics.

Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, and Yuanzhuo Wang. 2024. Unlocking the power of large language models for entity alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7566–7583, Bangkok, Thailand. Association for Computational Linguistics.

Bosung Kim, Taesuk Hong, Youngjoong Ko, and Jungyun Seo. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1737–1743, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Geoffrey N. Leech. 1974. *Semantics*. Penguin Books, Harmondsworth,.

Qingyang Li, Yanru Zhong, and Yuchu Qin. 2024. MoCoKGC: Momentum contrast entity encoding for knowledge graph completion. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14940–14952, Miami, Florida, USA. Association for Computational Linguistics.

Wei Li, Xuancheng Ren, Damai Dai, Yunfang Wu, Houfeng Wang, and Xu Sun. 2018. Sememe prediction: Learning semantic knowledge from unstructured textual wiki descriptions. *CoRR*, abs/1808.05437.

Jiashi Lin, Lifang Wang, Xinyu Lu, Zhongtian Hu, Wei Zhang, and Wenxuan Lu. 2024. Improving knowledge graph completion with structure-aware supervised contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13948–13959, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Zequn Sun, Guangyao Li, and Wei Hu. 2022. I know what you do not know: Knowledge graph embedding via co-distillation learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1329–1338, New York, NY, USA. Association for Computing Machinery.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Floyd G Lounsbury. 1956. A semantic analysis of the pawnee kinship usage. *Language*, 32(1):158–194.

John Lyons. 1968. *Introduction to theoretical linguistics*, volume 510. Cambridge university press.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164.

Michael Matuschek and Iryna Gurevych. 2014. High performance word sense alignment by joint modeling of sense distance and gloss similarity. In *Proceedings of COLING 2014, the 25th International*

Conference on Computational Linguistics: Technical Papers, pages 245–256, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024a. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350, Bangkok, Thailand. Association for Computational Linguistics.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024b. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 80 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A robust approach to aligning heterogeneous lexical resources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–478, Baltimore, Maryland. Association for Computational Linguistics.

Lucia Pitarch, Jordi Bernad, Lacramioara Dranca, Carlos Bobed Lisbona, and Jorge Gracia. 2023. No clues good clues: out of context lexical relation classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5607–5625, Toronto, Canada. Association for Computational Linguistics.

Fanchao Qi, Liang Chang, Maosong Sun, Sicong Ouyang, and Zhiyuan Liu. 2020. Towards building a multilingual sememe knowledge base: Predicting sememes for babelnet synsets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8624–8631.

Fanchao Qi, Yankai Lin, Maosong Sun, Hao Zhu, Ruobing Xie, and Zhiyuan Liu. 2018. Cross-lingual lexical sememe prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 358–368, Brussels, Belgium. Association for Computational Linguistics.

Fanchao Qi, Chuancheng Lv, Zhiyuan Liu, Xiaojun Meng, Maosong Sun, and Hai-Tao Zheng. 2022. Sememe prediction for BabelNet synsets using multilingual and multimodal information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Fanchao Qi, Ruobing Xie, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. Sememe knowledge computation: a review of recent advances in application and expansion of sememe knowledge bases. *Front. Comput. Sci.*, 15(5).

Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Qiang Dong, Maosong Sun, and Zhendong Dong. 2019. Openhownet: An open sememe-based lexical knowledge base. *CoRR*, abs/1901.09957.

Zile Qiao, Wei Ye, Dingyao Yu, Tong Mo, Weiping Li, and Shikun Zhang. 2023. Improving knowledge graph completion with generative hard negative mining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5866–5878, Toronto, Canada. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 23 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China. Association for Computational Linguistics.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA. PMLR.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*.

Alan Wake, Bei Chen, C. X. Lv, Chao Li, Chengen Huang, Chenglin Cai, Chujie Zheng, Daniel Cooper, Fan Zhou, Feng Hu, Guoyin Wang, Heng Ji, Howard Qiu, Jiangcheng Zhu, Jun Tian, Katherine Su, Lihuan Zhang, Liying Li, Ming Song, and 23 others. 2024. Yi-lightning technical report. *CoRR*, abs/2412.01253.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*, WWW '21, page 1737–1748, New York, NY, USA. Association for Computing Machinery.

Hansi Wang, Yue Wang, Qiliang Liang, and Yang Liu. 2025. LTRS: Improving word sense disambiguation via learning to rank senses. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1934–1942, Abu Dhabi, UAE. Association for Computational Linguistics.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. 2022a. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294, Dublin, Ireland. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Yue Wang, Qiliang Liang, Yaqi Yin, Hansi Wang, and Yang Liu. 2024. Disambiguate words like composing them: A morphology-informed approach to enhance Chinese word sense disambiguation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15354–15365, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3740–3751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Wierzbicka. 1996. *Semantics: Primes and Universals*. Oxford University Press.

Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, and Maosong Sun. 2017. Lexical sememe prediction via word embeddings and matrix factorization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4200–4206. AAAI Press.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *CoRR*, abs/1902.08564.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. 2021. Connect-the-dots: Bridging semantics between words and definitions via aligning word sense inventories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7741–7751, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Configuration for WSA

Table 10 presents the prompt template for WSA. This prompt guides LLMs to identify semantic consistency based on the information from both resources: parts-of-speech (PoS), sememe annotations, relations, and examples from HowNet, along with PoS, definitions, and examples from CCD.

我们需要识别HowNet和《现代汉语词典》(CCD) 之间语义一致的义项对。作为专家评估者，请你判断给定词的两个不同来源的义项之间的语义相似程度。(We need to identify semantically consistent sense pairs between HowNet and the Contemporary Chinese Dictionary. Please act as an expert evaluator to assess the semantic consistency between the two senses for the given word from the two sources.)

给定词{Word}，其在HowNet中的义项{HowNet_ID}，在CCD中的义项{CCD_ID}： (Given the word {Word}, its HowNet sense {HowNet_ID} and CCD sense {CCD_ID}:)

{HowNet_ID}信息：(The information of {HowNet_ID}:)
1. 词类信息：(PoS information:) {HowNet_PoS}
2. 义原标注信息：(Sememe annotation information:) {Sememe_Annotation}
3. 词义关系信息：(Lexico-semantic relation information:) {Relation}
4. 例句信息：(Example information:) {HowNet_Example}

{CCD_ID}信息： (The information of {CCD_ID}:)
1. 词类信息：(PoS information:) {CCD_PoS}
2. 释义信息：(Definition information:) {Definition}
3. 例句信息：(Example information:) {CCD_Example}

请评估这两个义项的语义相似程度：(Please assess the semantic consistency between the two senses:)
评分标准 [1-5]：(Rating scale [1-5]):
1: 非常低 - 两个义项描述的概念完全不同；(1: Very low - The two senses denote completely different concepts;)
2: 低 - 两个义项有微弱的一致性，但描述的概念明显不同； (2: Low - The two senses show weak consistency but denote obviously different concepts;)
3: 中等 - 两个义项有部分重叠，但也有明显差异； (3: Medium - The two senses have partial overlap but show significant differences;)
4: 高 - 两个义项描述的概念基本相同，只有细微差别； (4: High - The two senses basically denote the same concepts with subtle differences;)
5: 非常高 - 两个义项描述的是完全相同的概念。 (5: Very high - The two senses denote exactly the same concept.)

你的评分：(Your Rating Score:)

Table 10: Prompt template for identifying semantically consistent sense pairs between HowNet and CCD.

	human1	human2	human3	LLMs
human1	-	96.0	93.5	93.0
human2	96.0	-	95.5	94.0
human3	93.5	95.5	-	97.5
LLMs	93.0	94.0	97.5	-

Table 11: Inter-annotator agreement (%) metrics among LLMs and human annotators for WSA.

Each LLM is configured with a sampling temperature of 0.6 and a top-p value of 0.95 to generate three responses. We compute the mean of scores in these responses as the final score from each LLM.

B Human Evaluation for WSA

To validate the reliability of LLM-scored sense pairs, we randomly sample 200 ones and ask three linguistics researchers to perform binary judgments on their semantic consistency.

Table 11 shows that identifying semantically consistent sense pairs is a relatively straightforward task: there is substantial annotation consistency between LLMs and human annotators, as evidenced by comparable pairwise agreements (i.e., human-human: 0.935–0.960; human-LLM: 0.930–0.975). We also check samples where human annotators disagree with LLMs, finding that most errors are attributed to issues of sense granularity. For example, for the Chinese word "栓", HowNet defines it as "a switchable mechanism on a device", while CCD specifies "a switchable mechanism on a firearm". In this case, LLMs mistakenly categorize these two senses as semantically consistent. However, given the low frequency of such errors, their impact on overall performance is relatively limited.

C Configuration for Definition Generation

We use GPT-4o (OpenAI et al., 2024) to generate definitions for HowNet synsets with a temperature of 0.6 and a top-p value of 0.95. The prompt template is shown in Table 12.

D Details of HN7 and CWN5

For orthographic consistency with other Chinese datasets, we convert the data in CWN5 from traditional Chinese to simplified Chinese by OpenCC[7].

[7]https://github.com/BYVoid/OpenCC

给定同义词集{Synset_ID}，请你为该同义词集提供一个统一的释义描述，该释义应能够准确概括所有词语共同表达的核心语义。(Given the synset {Synset_ID}, please provide a unified definition that accurately summarizes the core semantics shared by all words in this synset.)

可利用的语义信息：(Available semantic information:)
1. 每个词的释义：(Definitions for each word:) {Definitions}
2. 义原标注信息：(Sememe annotation information:) {Sememe_Annotation}

释义生成原则：(Definition Generation Principles:)
1. 确保释义能覆盖同义词集中所有词语的共同语义，不包含仅属于个别词的特殊用法； (Ensure that the definition covers the common semantics of all words in the synset, without including special usages that belong only to some words;)
2. 释义应简洁明了，通常控制在10-20个汉字之间； (Ensure that the definition is concise, typically between 10-20 Chinese characters;)
3. 避免使用同义词集中已有的词语来解释自身。(Avoid using words in the synset to explain itself.)

你的释义描述：(Your Definition:)

Table 12: Prompt template for generating the definitions of synsets in HowNet.

Dataset	Antonymy	Hypernymy	Hyponymy	Holonymy	Meronymy	Material	Product	ALL
HN7	8,386	7,253	7,253	1,186	1,186	204	204	25,672
CWN5	1,103	1,974	1,974	172	172	N/A	N/A	5,395

Table 13: Statistics for the number of triples in HN7 and CWN5 across different relation types.

Dataset	Head Synset	Head Definition	Relation	Tail Synset	Tail Definition
HN7	{幼, 少 (*young*)}	年纪小的；未长成 (*young; immature*)	antonymy	{年迈, 老 (*old*)}	年岁大的；年长的 (*old; aged*)
	{教室 (*classroom*)}	学校内进行教学活动的房间 (*a room in a school where lessons take place*)	hypernymy	{屋子, 房间 (*room*)}	用墙隔开的供人居住或使用的建筑空间 (*a building space enclosed by walls for habitation or use*)
	{段, 段落 (*paragraph*)}	根据文章内容划分的相对独立的部分 (*relatively independent sections divided according to the content*)	holonymy	{正文 (*text*)}	著作的本文 (*the main body of a written work*)
	{建筑 (*building*)}	人工建造的供人们生产、生活的场所 (*man-made premises for production and living*)	material	{砖头、砖 (*brick*)}	用黏土烧制的长方形建筑材料 (*rectangular blocks of baked clay used for building*)
CWN5	{妇女, 女人 (*woman*)}	成年女子 (*a female adult*)	antonymy	{丁, 男人 (*man*)}	成年男子 (*a male adult*)
	{高中生 (*highschool student*)}	高级中学的学生 (*high school student*)	hypernymy	{生, 学生 (*student*)}	在学校内学习的人 (*a person who is studying at a school*)
	{叶子, 叶 (*leaf*)}	植物的营养器官之一，生于枝干上 (*one of the organs of vegetation, growing from stems*)	holonymy	{植物, 植 (*plant*)}	可自行制造养分，且不能运动的生物。 (*a living organism that can produce nutrients by its own, and lack the power of locomotion*)

Table 14: Examples from HN7 and CWN5 across different relation types.

Dataset	# Synset	# Relation	# Train	# Valid	# Test
WN18RR	40,943	11	86,835	3,034	3,134
HN7	10,939	4	13,624	1,702	1,703
CWN5	3,149	3	2,600	324	325

Table 15: Statistics of three benchmark datasets, WN18RR, HN7, and CWN5, including the number of synsets (# Synset), relation types (# Relation), and triples in training (# Train), validation (# Valid), and test (# Test) sets.

Hyperparameter	Value
epochs	50
learning rate	5e-5
optimizer	AdamW
task weight α	0.5
batch size	256
# negative samples	50
max # of tokens	64
soft prompt length	5

Table 16: The shared hyperparameters of the SP models for English and Chinese.

Table 13 shows the number of triples in HN7 and CWN5 across different relation types. Table 14 shows the examples from HN7 and CWN5 across different relation types.

E Acquisition of Sememe Labels

Formally, for the word sense definition d, the predicted sememe set $\hat{\mathcal{S}}_d$ is defined as:

$$\hat{\mathcal{S}}_d = \{s \in \mathcal{S} | p_s > \delta\}, \tag{11}$$

where p_s is the prediction score of s from $\boldsymbol{p}_{asp}$ as defined in Equation 5, and δ is the prediction score threshold.

The predicted main sememe $\hat{ms}_d$ is defined as:

$$\hat{ms}_d = \arg\max_{s \in \mathcal{S}} \boldsymbol{p}_{msp}[s], \tag{12}$$

where $\boldsymbol{p}_{msp}[s]$ denotes the MSP score of s.

To maintain prediction consistency between ASP and MSP tasks, we add $\hat{ms}_d$ into $\hat{\mathcal{S}}_d$. The value of δ is determined by the mean of F1-scores on ASP and MSP in the validation set of SememeDef, finally set to 0.45 for Chinese and 0.48 for English.

F Statistics for LP Benchmark Datasets

Table 15 shows the detailed statistics of the benchmark datasets. Following previous research, the reported number of triples does not include those with the removed relation types.

G Experimental Configuration

We adopt bert-base-uncased (Devlin et al., 2018) as the base model for English SP and LP, while using chinese-bert-base-wwm-ext (Cui et al., 2020) for Chinese. All of these BERT models consist of 12 layers with 768 hidden units.

For the SP task, we employ negative sampling during training to prevent excessive zero labels from affecting learning effectiveness. The hyperparameters are shown in Table 16. The models are evaluated every 250 steps on the validation set, with the best checkpoint[8] selected for LP enhancement.

For the LP task, we adopt consistent configurations across our models, setting the epochs to 30, learning rate to 5e-5, initial temperature τ to 0.05, InfoNCE margin γ to 0.02, max number of tokens to 64, with the mean pooling and AdamW (Loshchilov and Hutter, 2017) optimizer. The number of hidden layers for each MLP in the three-stage fusion module is set to 1, with the dimension set to double the input dimension. The model-specific hyperparameters for SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$ are aligned with SimKGC and MoCoKGC, respectively. Specifically, we first search for the best combination of hyperparameters for SimKGC and MocoKGC following their original papers. Then, SememeLP$_{\text{Sim}}$ and SememeLP$_{\text{MoCo}}$ adopt the same settings of overlapping hyperparameters as SimKGC and MoCoKGC, respectively. Detailed hyperparameters for SimKGC and SememeLP$_{\text{Sim}}$ are shown in Table 17, while those for MoCoKGC and SememeLP$_{\text{MoCo}}$ are presented in Table 18. The models are evaluated after each epoch on the validation set, with the best-performing checkpoint selected for the final evaluation on the test set. For the other baselines, the hyperparameters shared with SememeLP are aligned consistently to ensure a fair comparison, with the others following the settings described in their original papers.

All experiments are conducted with the deep learning framework PyTorch (Paszke et al., 2019) on a single NVIDIA A100 GPU (80GB memory).

H Relation Type Categorization in WN18RR

In Subsection 6.1, the relations in WN18RR are categorized to facilitate parallel comparison with the

[8]The best checkpoint is determined by the mean of MAP on ASP and F1 on MSP.

Hyperparameter	WN18RR	HN7	CWN5
batch size	1,024	1,024	512
negative sampling strategies	IB	IB, PB, SN	IB, PB, SN
weight for re-ranking	0.05	0.05	0.05
n-hop for re-ranking	5	3	3
pre-batch	0	2	2

Table 17: The shared hyperparameters of SimKGC and SememeLP$_{\text{Sim}}$ across different datasets. IB, PB, SN denote in-batch, pre-batch, and self-negatives sampling strategies, respectively.

Hyperparameter	WN18RR	HN7	CWN5
batch size	1,024	512	128
warmup	400	50	10
additional negatives	512	256	64
neighborhood sampling size	16	4	4
tail queue size	16,384	4,000	750
relation prompt length	4	4	4

Table 18: The shared hyperparameters of MoCoKGC and SememeLP$_{\text{MoCo}}$ across different datasets.

Relation in WN18RR	Relation	# Triple
hypernym, instance_hypernym	hypernymy	1,373
hypernym^{-1}, instance_hypernym^{-1}	hyponymy	1,373
has_part^{-1}, member_meronym^{-1}	holonymy	425
has_part, member_meronym	meronymy	425
derivationally_related_form, verb_group, similar_to, synset_domain_topic_of, member_of_domain_usage, member_of_domain_region, also_see, and their inverse relations	others	2,672

Table 19: Mapping of relations in WN18RR to the relation types. The notation $^{-1}$ denotes the inverse operation of a relation type.

other two datasets. Detailed mapping information is provided in Table 19.

I Case Studies

SememeLP achieves significant performance by incorporating sememe knowledge. To better understand the value of such knowledge for LP, we further illustrate this through case studies. Examples of predictions from MocoKGC and SememeLP$_{\text{MoCo}}$ are shown in Table 20. From them, we have the following observations:

(1) Lexico-semantic relations are reflected by the differences of sememes. For example, the antonymy relation between 废品 (*unqualified artificial product*) and 正品 (*qualified artificial product*) is reflected by `unqualified`|不合格 and `qualified`|合格; the hypernym-hyponym relation between 容器 (*a utensil used for holding articles*) and 碟子 (*a small shallow vessel for holding food*) is reflected by `artifact`|人工物 and `edible`|食物; The holonym-meronym relation between *Germany* and *Bonn* is reflected by `country`|国家 and `city`|市. By incorporating sememe knowledge, SememeLP$_{\text{MoCo}}$ makes more reasonable predictions in these cases than MocoKGC;

(2) The finite set of sememes enables these differences to be generalized feasibly. For example, in the training set of WN18RR, synsets within holonym-meronym relations, such as {*Italy*}-{*Rome*} and {*China*}-{*Hangzhou*} show systematic differences in their sememes (i.e. `country`|国家 and `city`|市). SememeLP$_{\text{MoCo}}$ can thus transfer this mapping to the meronym prediction for {*Germany*}, and provide a more reasonable answer than MoCoKGC;

(3) Sememe information highlights the core elements of word senses, which helps to alleviate the problems of definitions. For instance, MoCoKGC incorrectly predicts {*Volgograd*} as a part of {*Germany*}, possibly due to the appearance of the word *German* in its definition *a city in the European part of Russia on the Volga; site of German defeat in World War II in the winter of 1942-43*.

However, SememeLP fails to achieve performance gains on some subsets of the datasets. We further conduct an error analysis to discuss these cases.

For the hypernymy and hyponymy subsets of CWN5, we observe that SememeLP$_{\text{MoCo}}$ makes some reasonable but incorrect predictions due to the inconsistencies of taxonomy between CWN and HowNet. As shown in Table 21, 教授 (*a teacher who teaches at universities*) is categorized as a type of *occupation* in CWN, while it is viewed as a type of *human* in HowNet. This stems from different sense facet perspectives. It also indicates that con-

({*Germany*}, meronymy, ?) from WN18RR	
Head	{***Germany***} **Definition**: a republic in central Europe; split into East Germany and West Germany after World War II and reunited in 1990 **Sememes**: {place|地 方, Europe|欧 洲, country|国 家, politics|政, Germany|德 国, properName|专}
Relation	**meronymy**
Ground-Truth	{***Bonn***} **Definition**: a city in western Germany on the Rhine River; was the capital of West Germany between 1949 and 1989 **Sememes**: {place|地方, city|市, Germany|德国, properName|专}
MocoKGC	{***Volgograd***} **Definition**: a city in the European part of Russia on the Volga; site of German defeat in World War II in the winter of 1942-43 **Sememes**: {place|地方, city|市, Russia|俄罗斯, properName|专}
SememeLP$_{\text{MoCo}}$	{***Wiesbaden***} **Definition**: a city in western Germany; a spa since Roman times **Sememes**: {place|地方, city|市, Germany|德国, properName|专}
({废品}, antonymy, ?) from HN7	
Head	{废品 (***waste product***)} **Definition**: 不合格的人工制品 (*unqualified artificial product*) **Sememes**: {artifact|人工物, unqualified|不合格})
Relation	**antonymy**
Ground-Truth	{正品 (***qualified product***)} **Definition**: 符合标准的人工制品 (*qualified artificial product*) **Sememes**: {artifact|人工物, qualified|合格})
MocoKGC	{精品 (***refined product***)} **Definition**: 高品质的人工制品 (*high-quality artificial product*) **Sememes**: {artifact|人工物, refined|精})
SememeLP$_{\text{MoCo}}$	正品 (***qualified product***)} **Definition**: 符合标准的人工制品 (*qualified artificial product*) **Sememes**: {artifact|人工物, qualified|合格})
({碟子}, hypernymy, ?) from CWN5	
Head	{碟子 (***small plate***)} **Definition**: 盛放食物的小浅底器皿 (*a small shallow vessel for holding food*) **Sememes**: {tool|用具, put|放置, edible|食物}
Relation	**hypernymy**
Ground-Truth	{容器 (***container***)} **Definition**: 用来盛装物品的器具 (*a utensil used for holding articles*) **Sememes**: {tool|用具, put|放置, artifact|人工物}
MocoKGC	{杯, 杯子 (***cup***)} **Definition**: 饮用的柱状器皿 (*a cylindrical vessel for drinking*) **Sememes**: {tool|用具, put|放置, drink|喝, drinks|饮品}
SememeLP$_{\text{MoCo}}$	{容器 (***container***)} **Definition**: 用来盛装物品的器具 (*a utensil used for holding articles*) **Sememes**: {tool|用具, put|放置, artifact|人工物}

Table 20: Examples of top-1 prediction results from MocoKGC and SememeLP$_{\text{MoCo}}$ across different datasets. The sememe information is predicted by our sememe knowledge encoder, with the first sememe denoting the main sememe for the word sense.

({*waterfall*}, holonymy, ?) from WN18RR	
Head	{***waterfall***} **Definition**: a steep descent of the water of a river **Sememes**: {waters|水域, flow|流}
Relation	**holonymy**
Ground-Truth	{***river***} **Definition**: a large natural stream of water (larger than a creek) **Sememes**: {waters|水域, linear|线}
SememeLP$_{\text{MoCo}}$	{***watercourse***} **Definition**: a natural body of running water flowing on or under the earth **Sememes**: {waters|水域, flow|流, water|水}
({教授}, hypernymy, ?) from CWN5	
Head	{教授 (***professor***)} **Definition**: 任教于大学的老师 (*a teacher who teaches at universities*) **Sememes**: {human|人, occupation|职位, teach|教, education|教育}
Relation	**hypernymy**
Ground-Truth	{行, 职业, 工作 (***occupation***)} **Definition**: 在社会中所担任的职务 (*a role taken in society*) **Sememes**: {occupation|职位, affairs|事务, human|人, status|身分}
SememeLP$_{\text{MoCo}}$	{员, 人 (***employee***)} **Definition**: 从事特定工作的人 (*a person who performs a specific job*) **Sememes**: {human|人, occupation|职位, employ|雇用}

Table 21: Examples of incorrect top-1 predictions from SememeLP$_{\text{MoCo}}$ across different datasets. The sememe information is predicted by our sememe knowledge encoder, with the first sememe denoting the main sememe for the word sense.

	Dataset	Training Params	Sememe Knowledge Params	Training Time per Epoch	MRR (%)
MoCoKGC	WN18RR	424M	0M	8.6 min	74.2
SememeLP$_{\text{MoCo}}$	WN18RR	552M	63M	8.9 min	75.1
MoCoKGC	HN7	402M	0M	1.0 min	76.5
SememeLP$_{\text{MoCo}}$	HN7	530M	17M	1.0 min	80.5

Table 22: Comparisons of model efficiency between MoCoKGC and SememeLP$_{\text{MoCo}}$.

ventional evaluation metrics may not appropriately assess the results in such cases.

For the holonymy subset of both WN18RR and HN7, SememeLP$_{\text{MoCo}}$ underperforms due to inaccuracies of SP. Table 21 shows that for {*waterfall*}, SP fails to identify water|水 as the ground-truth main sememe. This prevents SememeLP$_{\text{MoCo}}$ from leveraging the meronym-holonym relation between water|水 and waters|水域 for correct prediction.

J Cost-Benefit Analysis

We quantify and compare the computational costs of SememeLP and MoCoKGC under the same experimental condition (NVIDIA A100 GPU), as shown in Table 22.

Due to employing additional fusion modules, the training parameters of our method increase by 128M. Additionally, SememeLP requires extra memory consumption (63M for WN18RR and 17M for HN7) to store the sememe features of nodes in KGs. However, the training time per epoch does not increase significantly (8.6min vs. 8.9 min for WN18RR and 1.0min vs. 1.0 min for HN7), and SememeLP achieves significant improvements of 0.9 and 4.0 MRR points on WN18RR and HN7, respectively. In knowledge-intensive NLP tasks, it is worthwhile to pay such a relatively limited computational investment for introducing sememe knowledge to benefit LP.

It should also be noted that our method requires fine-tuning an additional sememe knowledge encoder, with quantifiable training time: 2.5 min/epoch $\times$ 50 epochs for English and 1.6 min/epoch $\times$ 50 epochs for Chinese. However, this is a once-for-all effort and is inappropriate to be directly counted as the fine-tuning cost for a specific

LP dataset. We also find that introducing sememe knowledge helps to accelerate convergence, as supported by validation performance trends at training time, where SememeLP achieves the best performance at the 14th epoch on WN18RR and 18th epoch on HN7 compared to MoCoKGC's 22nd and 23rd epoch. This offsets the time cost of training the sememe knowledge encoder to a certain extent.