

Article

A Sememe Prediction Method Based on the Central Word of a Semantic Field

Guanran Luo  and Yunpeng Cui 

Agriculture Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; 82101212469@caas.cn

* Correspondence: cuiyunpeng@caas.cn

Abstract: A “sememe” is an indivisible minimal unit of meaning in linguistics. Manually annotating sememes in words requires a significant amount of time, so automated sememe prediction is often used to improve efficiency. Semantic fields serve as crucial mediators connecting the semantics between words. This paper proposes an unsupervised method for sememe prediction based on the common semantics between words and semantic fields. In comparison to methods based on word vectors, this approach demonstrates a superior ability to align the semantics of words and sememes. We construct various types of semantic fields through ChatGPT and design a semantic field selection strategy to adapt to different scenario requirements. Subsequently, following the order of word–sense–sememe, we decompose the process of calculating the semantic sememe similarity between semantic fields and target words. Finally, we select the word with the highest average semantic sememe similarity as the central word of the semantic field, using its semantic primes as the predicted result. On the BabelSememe dataset constructed based on the sememe knowledge base HowNet, the method of semantic field central word (SFCW) achieved the best results for both unstructured and structured sememe prediction tasks, demonstrating the effectiveness of this approach. Additionally, we conducted qualitative and quantitative analyses on the sememe structure of the central word.

Keywords: sememe prediction; semantic field; HowNet

Citation: Luo, G.; Cui, Y. A Sememe Prediction Method Based on the Central Word of a Semantic Field. *Electronics* **2024**, *13*, 413. <https://doi.org/10.3390/electronics13020413>

Academic Editor: Arkaitz Zubiaga

Received: 9 December 2023

Revised: 7 January 2024

Accepted: 16 January 2024

Published: 19 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to linguistic definitions, a “sememe” is considered the smallest indivisible unit of meaning [1]. One of the most well-known databases of sememes is HowNet [2], which has been constructed over the past 30 years and currently comprises 2540 sememes, annotating over 230,000 Chinese and English words.

Figure 1 illustrates the sememe annotation for the word “husband” in HowNet, including multiple senses, such as “丈夫 (husband)” and “节约 (husband)”, each of which correspond to a definition of the word and are further annotated with a sememe tree. The sememe tree is a set of structured sememes used to express the potential semantics of a word sense [2]. Nodes in a sememe tree, such as “human | 人,” “family | 家庭,” and “male | 男,” represent sememes, while edges, such as “belong” and “modifier”, denote the dynamic roles between sememes.

HowNet can be applied to various natural language processing (NLP) tasks, including semantic representation learning [3], word similarity calculations [4–6], word sense disambiguation [7,8], sentiment analysis [9,10], web text mining [11,12], knowledge graph construction [13], defense against textual backdoor attacks [14], and improving text readability [15], among others. Compared to other knowledge bases, HowNet has the following advantages:

(1) In the field of natural language understanding, sememes are closer to the semantic essence. Unlike word-level knowledge bases, such as WordNet, a sememe knowledge base, such as HowNet, surpasses the barrier of individual words, serving as a crucial pathway for a deep understanding of the semantic information behind words. Existing works on sememe-based neural language models [16–21] leverage this characteristic of sememes.

(2) Sememe knowledge bases can be more effectively integrated into deep learning. HowNet focuses on the semantic composition relationship from sememes to words. A word often corresponds to a processing unit in existing deep learning models, and the sememes of that word can be directly introduced as semantic labels into the corresponding processing unit—a capability that other word-level knowledge bases cannot achieve [17,18].

(3) In low-data scenarios, sememe knowledge bases help address the issue of poor performance in deep learning models. Due to the limited number of sememes, they can be adequately trained, and well-trained sememes can alleviate the problem of insufficient model training in low-data situations. This characteristic is most notably applied in word representation learning, where the introduction of sememes can enhance the performance of word vector methods for low-frequency words [3].

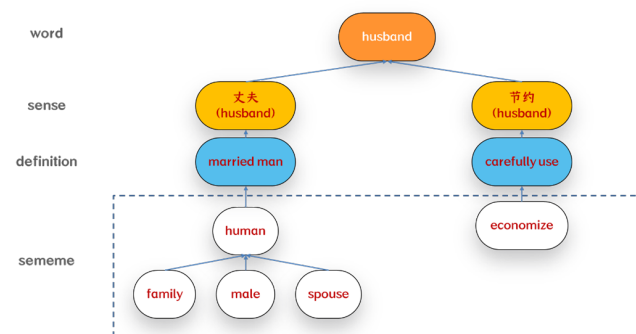


Figure 1. Sememe annotations of the word “husband” in HowNet. Each sense corresponds to a distinct definition of the word and is further annotated with a set of sememes, which serve to express its underlying semantics.

However, as a manually constructed semantic knowledge base, HowNet faces challenges regarding its high extension and update costs, slow speed, and difficulties in adapting to the continuous emergence of new words and rapid changes in word meanings. To address this issue, researchers have attempted to improve sememe annotation efficiency through automated sememe prediction [22]. Previous research has mainly focused on supervised methods, such as collaborative filtering-based methods, including sememe prediction with word embeddings (SPWE), and matrix factorization-based methods, such as sememe prediction with sememe embeddings (SPSE) [22]. Later approaches enhanced prediction accuracy by incorporating additional semantic information during training. Jin [23], Lyu [24], and Kang [25], among others, utilized the character information of Chinese characters during prediction tasks. Bai [26] and Du [27] introduced dictionary definitions, and Qi [28] integrated word embeddings from other languages.

Aligning the semantics of a word in both word vector space and sememe space is challenging because these methods depend on high-quality word vectors. The difficulty arises from sememe annotation relying on expert knowledge instead of textual co-occurrence information. When mapping sememe vectors and word vectors to the same semantic space, there is an inevitable loss of semantic information.

As shown in Table 1, the words “water” and “bottled water”, “mineral water”, and “hard water” share certain semantics, often co-occurring in texts. Therefore, they exhibit closer distances in the Word2Vec word vector space. However, their annotations in HowNet show strong subjectivity, potentially resulting in discrepancies between Word2Vec similarity and sememe similarity. For instance, words such as “water” and “mineral water” may have a high Word2Vec similarity but a lower sememe similarity, while words such as “water” and “hard water” may have identical sememes but a lower Word2Vec similarity.

Table 1. Word pairs similarity in different semantic spaces.

Word Pairs	Word2vec Similarity	Sememe Similarity
Water, Mineral Water	0.73	0.54
Water, Bottled Water	0.53	0.54
Water, Hard Water	0.71	1
Bottled Water, Mineral Water	0.83	1
Bottled Water, Hard Water	0.56	0.54
Mineral Water, Hard Water	0.65	0.54

To address these issues, we propose an unsupervised method based on semantic field central word (SFCW) to predict sememes for an unlabeled new word. A semantic field is a collective of interconnected, mutually restrictive, distinctive, and interdependent terms in terms of semantics. The semantic scope encapsulated by this collective is referred to as the field of this semantic field. Words at the center of the field embody shared semantic elements of all terms, symbolizing the core concept of the semantic field. If a semantic field is constructed around a particular word, the semantics of the central word should be very close to that word.

Specifically, we first construct two types of semantic fields for words using a large language model, namely, taxonomic fields composed of hypernyms and synonymous semantic fields composed of synonyms or near-synonyms. For word groups consisting of multiple words, we create a special semantic field composed of the characters obtained by splitting the words. We then quantitatively analyze the sememe fitness between these three types of semantic fields and the sememe fitness between words. Based on the analysis results, we formulate a strategy for selecting semantic fields. Finally, we use a mixed method that considers the sememe tree structure to calculate sememe similarity and obtain the central word of the semantic field. We use the sememes of the central word as the prediction results. On the BabelSememe dataset, SFCW outperforms all other baseline models, achieving the best results in both unstructured and structured sememe prediction.

In summary, this paper makes the following contributions:

- We construct three types of semantic fields to match words and sememes and design a semantic field selection strategy to expand the applicability of semantic fields and improve the accuracy of sememes.
- By computing the semantic field's central word for prediction, a better alignment between words and sememes is achieved, eliminating the need for vector training.
- The proposed SFCW achieves the best unstructured and structured sememe prediction results on the publicly available BabelSememe dataset. Additionally, we further qualitatively and quantitatively analyze the sememe structure of the central word.

2. Related Work

2.1. HowNet

HowNet describes the meanings of all words in terms of the smallest semantic units called “sememes.” A word in HowNet may have one or more senses, each composed of structured sememes. HowNet defines all senses in the format shown in Table 2.

For example, “human”, although it is a highly complex concept representing an aggregate of various attributes, can also be viewed as a sememe. We envision that all concepts can be decomposed into diverse sememes. Simultaneously, we imagine the existence of a finite set of sememes, where these sememes combine to form an infinite set of concepts. If we can grasp this limited set of sememes and utilize it to describe relationships between concepts as well as relationships between attributes, we may have the opportunity to establish the envisioned knowledge system.

Despite the tremendous potential of HowNet as a knowledge resource, its creators acknowledge that the most challenging aspects of HowNet research and construction lie in determining the primary and secondary features of words, organizing them, and

annotating meanings for each word one by one [2]. These challenges are crucial factors hindering the development, promotion, and application of HowNet itself.

Table 2. The knowledge dictionary is the foundational document of the HowNet system. In this document, each word is associated with its sense and description, forming a record. Each record in the knowledge dictionary for a specific language primarily consists of four components: W_X = word, G_X = part-of-speech, E_X = example sentence, and DEF = definition, where X represents the language type, with C for Chinese and E for English.

Sense	HowNet Dictionary Record
医师 doctor	NO. = 000000064959 W_C = 医师 G_C = noun E_C = W_E = doctor G_E = noun E_E = DEF = {human 人:HostOf = {Occupation 职位}, domain = {medical 医}, {doctor 医治:agent = {~}}
患者 patient	NO. = 000000130638 W_C = 患者 G_C = noun E_C = W_E = patient G_E = noun E_E = DEF = {human 人:domain = {medical 医}, {SufferFrom 罹患:experiencer = {~}}, {doctor 医治:patient = {~}}

2.2. Sememe Prediction

Research on sememe prediction can be divided into two stages: (1) unstructured sememe prediction and (2) structured sememe prediction.

Initially, researchers simplified the task of sememe prediction by only predicting a sememe's set of words and ignoring the hierarchical structure of sememes. Xie et al. [22] proposed two simple yet effective methods: the collaborative filtering-based method, SPWE, and the matrix factorization-based method, SPSE. Bai [26] and Du [27] introduced dictionary definitions into sememe prediction, enabling the selection of more accurate sememes for words with given definitions. Jin [23], Lyu [24], and Kang [25] improved sememe prediction models by incorporating the character information of Chinese characters. Wang et al. [29] utilized the expert-annotated knowledge base CilinE for sememe prediction, achieving better performance. Qi [28] proposed cross-language vocabulary sememe prediction, aligning word embeddings between languages covered by HowNet and new languages to predict sememes for words in the new language. Qi et al. constructed the multilingual BabelSememe [18] dataset based on BabelNet [30], aligning the definitions of BabelNet synsets with sememes in HowNet.

Du et al. [27] were the first to incorporate structured sememe evaluation into sememe prediction research. They proposed a simple method for constructing sememe trees using bidirectional long short-term memory (BiLSTM) to predict the sememe set for each word. Once each word had a set of sememes, they used a multi-layer neural network to predict edges between sememes, gradually removing low-confidence edges until there were no cycles. However, their evaluation only involved dividing each sememe's hierarchical structure into a series of triplets consisting of "node–edge–node" and used the similarity between triplets to assess the local structural similarity of sememe trees.

Subsequently, Liu et al. [31] introduced the task of sememe tree prediction and introduced depth-first generation methods and recommendation-based methods for sememe tree prediction. Ye et al. [32] designed a Transformer-based sememe tree generation model, improving attention mechanisms and achieving better results. Additionally, they intro-

duced the strict F1-metric to measure the overall structural similarity between sememe trees, further enhancing the evaluation metrics for structured sememes.

3. Methods

As illustrated in Figure 2, our method consists of three main steps: semantic field organization (Section 3.1), semantic field selection (Section 3.2), and central word selection (Section 3.3). Initially, specific information is inputted into ChatGPT to generate three types of semantic fields. Subsequently, an optimal semantic field is chosen using a devised strategy. Within the selected semantic field, the similarity is computed in the order of word–sense–sememe, and the word with the highest average similarity is identified as the central word. The sememes of this central word serve as the prediction result.

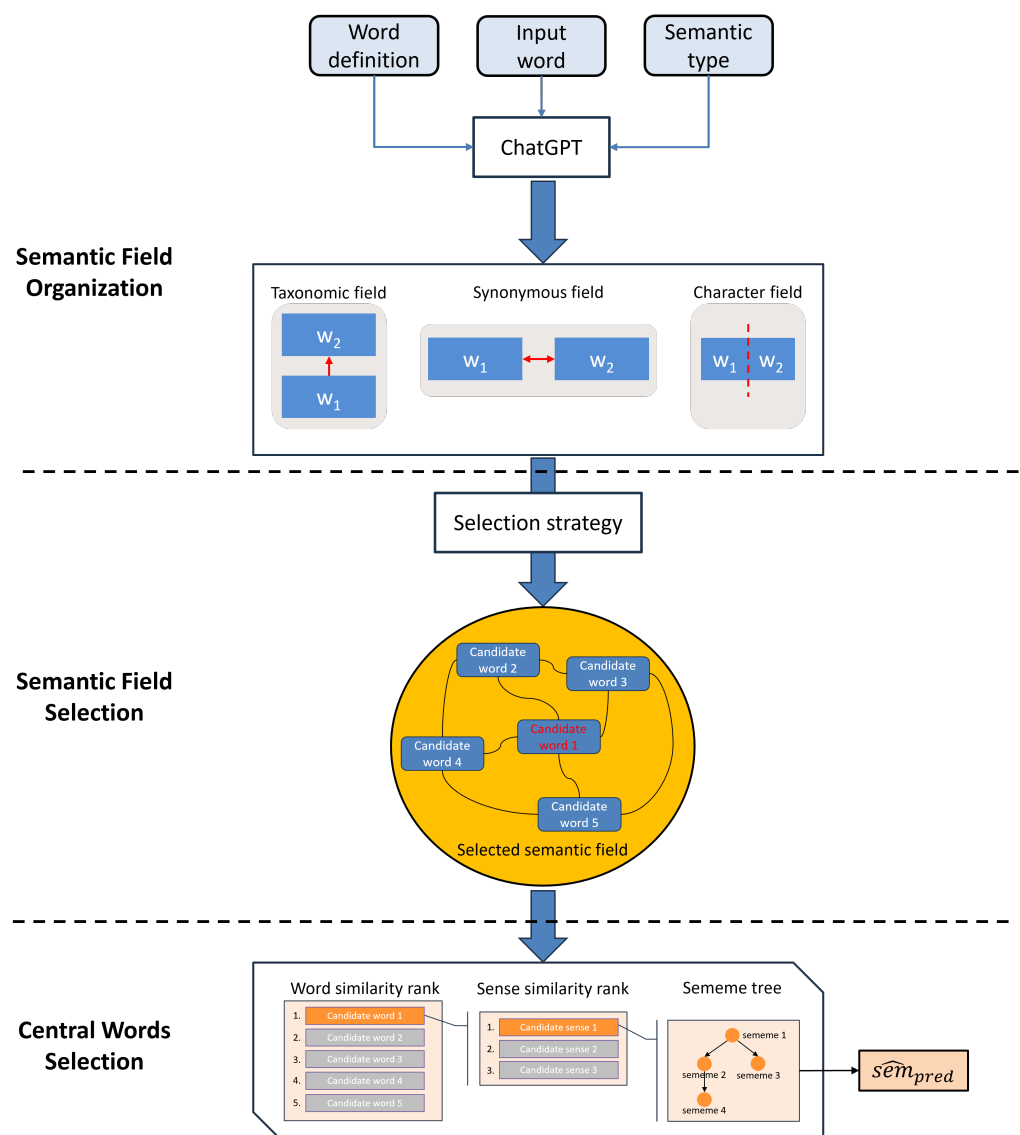


Figure 2. The primary workflow of our method.

3.1. Semantic Field Organization

While semantic knowledge bases, such as WordNet and BabelNet, offer synonym sets and hypernym paths for constructing semantic fields, their coverage is limited. Not all words have readily available synonyms or hypernyms.

Therefore, we used ChatGPT to generate words with specific semantic relationships to build semantic fields. These fields included taxonomic fields composed of hierarchical relationships and synonymous fields composed of synonymous or near-synonymous

relationships. Additionally, we divided multi-word expressions into characters to form a special type of semantic field. We constructed these three types of semantic fields because we believed they have stronger semantic associations with words than other types of semantic fields.

First, we used ChatGPT to create corresponding prompt templates for different types of semantic fields. We constructed the following prompt:

- As a specialist in the field of natural language processing, give #type of #word (#definition), separated by commas, no capitalization, no number, no commentary.

This template specifies that the model's role is that of an expert in the field of natural language processing. The task is to generate synonyms or hypernyms and format the response according to specific text processing rules. Here, #type represents the semantic relationship based on which the semantic field is constructed, #word represents the word around which the semantic field is built, and #definition corresponds to the definition of that word in BabelNet. Including the definition in the prompt helps eliminate ambiguity. Additionally, we included qualifiers in the task instruction, such as "separated by commas, no capitalization," to ensure that the generated answers followed a fixed format, facilitating subsequent data processing. Table 3 displays an example of the prompt we constructed.

Table 3. Example of a prompt to build a semantic field.

Prompt	Type	Word	Definition
{ "role": "user", "content": "As a specialist in the field of natural language processing, give synonyms of waffle (pancake batter baked in a waffle iron), separated by commas, no capitalization, no number, no commentary", } { "role": "assistant", "content": "waffle: battercake, griddlecake, pancake, hotcake, flannel cake"} }	synonyms	waffle	pancake batter baked in a waffle iron

3.2. Semantic Field Selection

When examining the three types of semantic fields constructed in the preceding section as a collective entity and selecting a word as the central focus, two key issues emerged. Firstly, diverse semantic field categories exhibit distinct semantic tendencies, potentially introducing interference during central word selection. Secondly, the semantics of the chosen central word may incline towards favoring semantic fields with a greater number of words rather than accurately representing the central concept of the semantic field.

To address these issues, in this Section, we conducted confirmatory experiments on the fit between different types of semantic fields and the sememes of words to determine the best strategy for selecting semantic fields. This strategy takes into account the applicable scenarios of different types of semantic fields and aims to narrow down the calculation scope of the central word.

In this experiment, we created a set of sememes for each word in the semantic field corresponding to all senses of the word and compared it with the set of sememes of the unlabeled target word. If we consider the sememes of words as label categories, then the comparison between these two sets can be seen as a multi-label classification task, using the F1 score as a metric to measure the similarity between sets of sememes.

We conducted experiments using the BabelSememe [18] dataset. Some words in the BabelSememe dataset were originally not defined in HowNet, and their sememes were obtained through automatic prediction. We refer to these words as OOV (out-of-vocabulary) words, while the others are considered IV (in-vocabulary) words. We randomly selected 5461 IV words, representing about one-third of the total data. To maintain a consistent ratio with the overall data distribution, we chose 545 OOV words since the OOV-to-IV word ratio in BabelSememe is approximately 1:10. During the selection process, we used the OpenHowNet [33] API to check if words in BabelSememe had sememe definitions. If the library returned an empty list, we considered it an OOV word; otherwise, it was an IV word.

For each selected word, we generated three semantic fields. The experimental results in Figure 3 show that:

1. Character fields perform best for in-vocabulary words, with a significant decrease in performance for out-of-vocabulary words. Additionally, character fields are only suitable for phrases composed of multiple words. This is because, for a single word, the character domain is just the word itself. Predicting the word using itself is not sensible.
2. Synonymous fields have overall stable performance, with good performance for both in-vocabulary and out-of-vocabulary words. However, large models may not generate correct synonyms for all words, especially proper nouns (names, locations).
3. Taxonomic fields have the lowest sememe compatibility, but their advantage lies in that almost all words can find corresponding hypernyms, making them suitable for a wide range of scenarios.

Based on these conclusions, we formulated the following semantic field selection strategy as shown in Figure 4: first, we searched for the central word in the synonymous field. If the synonymous field for that word was an empty set, then we proceeded to search for the best sense in the character field. If the character field was also an empty set, then we finally searched for the central word in the taxonomic field.

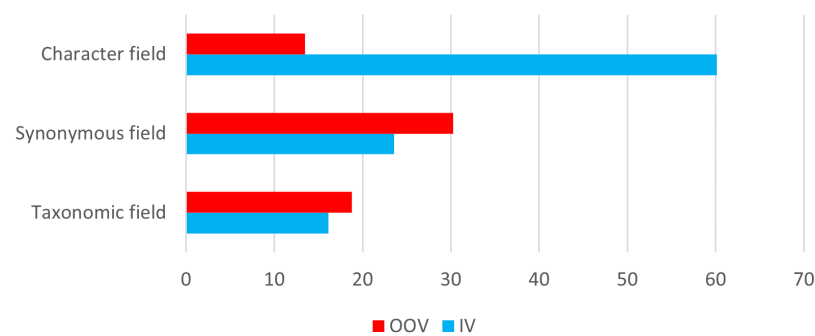


Figure 3. Comparison of F1 scores between OOV (out-of-vocabulary) words and IV (in-vocabulary) words.

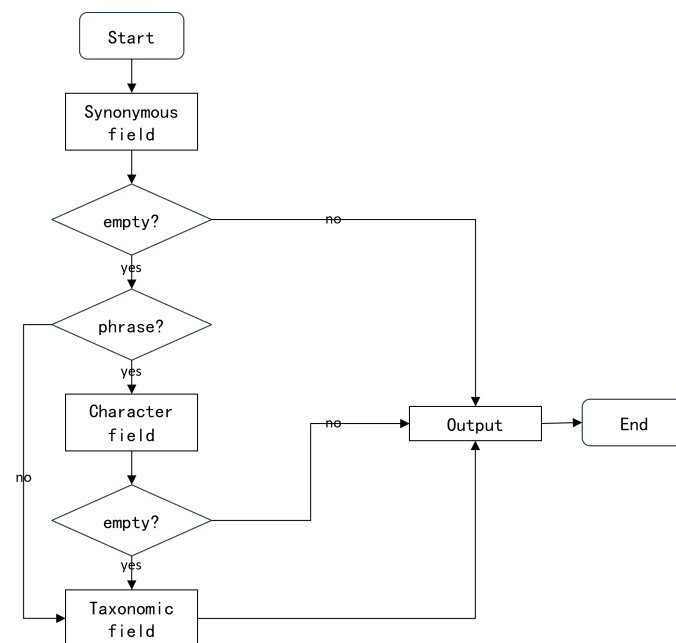


Figure 4. Selection process of semantic fields.

3.3. Central Word Selection

3.3.1. Semantic Relevance of Senses

In HowNet, a word encompasses multiple senses, each associated with a distinct semantic meaning. Consequently, when computing the central word, it becomes imperative to ascertain the sense of the word that pertains to the semantic field. Simply put, the central concept of the semantic field hinges on a specific definition corresponding to each word within that field.

We assume that within a semantic field, there are two words, w_1 and w_2 , each with sense sets Q_1 and Q_2 , respectively. If w_1 and w_2 share a specific semantic relationship, then there should be at least one pair of senses in Q_1 and Q_2 that share the same semantic relationship. This is because the definition of a sense is a subset of the semantic meaning of a word.

As shown in Figure 5, the candidate words of “water meter” include “instrument,” “water gauge,” and “meter.” Senses of the same word may have significantly different definitions. For example, the sense of “instrument” includes “仪器 (instrument): a device that requires skill for proper use,” “爪牙 (instrument): a person used by another to gain an end,” etc. Clearly, not all of these senses belong to the semantic fields of their respective words. For the sense “仪器 (instrument)”, we can find the semantically related senses “水尺 (water gauge)” and “仪表 (meter)” in the other two words, while there is no semantically related sense for the sense “爪牙 (instrument).” Therefore, the sense “仪器 (instrument)” should belong to the semantic fields of its respective words.

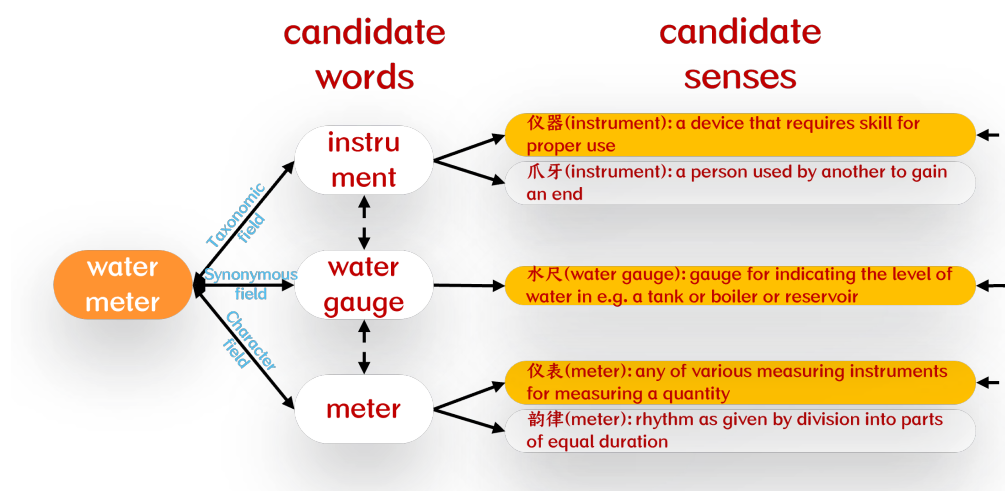


Figure 5. The candidate words form the context of the target word. If the candidate words are semantically related to each other, then some of their senses are also semantically related.

Due to the significant semantic differences among senses of a word, when computing the central word, we need to know not only which word is the central word but also which sense of the central word corresponds to the semantic field. In HowNet, we can address these two issues by calculating the sememe similarity, as sememe similarity not only measures the semantic relevance between words but also further assesses the membership of senses within the semantic field.

Next, we will introduce how to calculate sememe similarity between words and sense-to-sense sememe similarity in a top-down manner.

3.3.2. Sememe Similarity between Words

Fan et al. [4] argued that the similarity between words is determined by the maximum similarity between their respective sense components. However, they overlooked the fact that in different contexts, the maximum similarity of sense components cannot be used as the similarity between two words. Here, we further supplement Liu’s definition of sememe similarity between words: in the same semantic field, for two words w_1 and w_2 , if

w_1 has n sense components such that $S_{11}, S_{12}, \dots, S_{1n}$, and w_2 has m sense components such that $S_{21}, S_{22}, \dots, S_{2m}$, then the similarity between w_1 and w_2 is determined by the maximum similarity between their respective sense components. Formally, we define:

$$\text{sim}(w_1, w_2) = \max_{i=1 \dots n, j=1 \dots m} \text{sim}(S_{1i}, S_{2j}) \quad (1)$$

This approach transforms the calculation of word sememe similarity into the calculation of similarity between their respective sense components.

3.3.3. Sememe Similarity between Senses

This article employs a calculation method based on the path length between sememes. Assuming the distance between two sememes is denoted as d , it is defined as follows:

$$\text{sim}(\text{sem1}, \text{sem2}) = \frac{\mu}{d + \mu} \quad (2)$$

where sem1 and sem2 represent two sememes, d is the path length of the sememe in the hierarchical system, and μ is a tunable parameter.

To compute the sememe similarity between entries, we adopted the hybrid method proposed by Li [6], which considers both the Markov model and structural similarity of sememe trees. The calculation method is as follows:

$$\text{sim}(o_1, o_2) = \text{markovSim}(o_1, o_2) \times \alpha + \text{treeSim}(o_1, o_2) \times \beta \quad (3)$$

where the hyperparameters α and β determine the weights of these two similarity calculation methods. markovSim is obtained by computing the transition probabilities between sememe nodes using the Markov model. Initially, the sememe similarity of the main node s_1 is computed, followed by the average sememe similarity of the child nodes s_2 . Different weights are assigned to the main node and its child nodes:

$$\text{markovSim} = s_1 \times m + s_2 \times n \quad (4)$$

where $m > n$ and $m + n = 1$. The calculation of s_2 is recursive, resulting in the overall sememe tree's similarity. On the other hand, treeSim assigns weights based on the depth of nodes in the sememe tree:

$$\text{weight}_{\text{sememe}} = \left(\frac{1}{1 + d_{\text{tree}}} \right)^l \quad (5)$$

where d_{tree} represents the depth of the sememe in the sememe tree, and l denotes the level at which the sememe is located.

treeSim includes both the node similarity and edge similarity:

$$\text{nodeSim}(t1, t2) = \frac{2 * \sum \text{weight}(m_{\text{node}})}{\sum \text{weight}(n_{t1}) + \sum \text{weight}(n_{t2})} \quad (6)$$

$$\text{edgeSim}(t1, t2) = \frac{\sum e_{t1} \times \sum m_{\text{edge}} + \sum e_{t2} \times \sum m_{\text{edge}}}{2 \times \sum e_{t1} \times \sum e_{t2}} \quad (7)$$

Finally, the structural similarity of the sememe tree can be expressed as:

$$\text{treeSim}(t1, t2) = \varphi \times \text{nodeSim}(t1, t2) + \delta \times \text{edgeSim}(t1, t2) \quad (8)$$

where φ and δ are hyperparameters and $\varphi + \delta = 1$.

3.3.4. Word Filtering

While using ChatGPT instead of a knowledge base to construct a semantic field can increase the coverage of words, it introduces the issue of “illusion”, where the generated words by ChatGPT may not necessarily conform to the semantic relationships we have set, thus interfering with the central semantic meaning of the field.

To mitigate the impact of such noise words on the selection of the central word, we first calculated the sememe similarity for all words in the semantic field. Words with sememe similarity below a predefined threshold were filtered out, as this implies a significant deviation of all senses of that word from the central semantic meaning of the field.

Subsequently, we recalculated the sememe similarity on the semantic field, now devoid of filtered-out noise words. The word with the highest average sememe similarity was then chosen as the central word, with its senses serving as the central semantic meaning.

4. Experiments

4.1. Dataset

We utilized the BabelSememe dataset constructed by Qi [18]. Ye [32] also employed this dataset to evaluate their structured sememe prediction model. The format of the BabelSememe dataset is shown in Table 4. We chose this dataset because it aligns English synonyms from BabelNet with senses from HowNet, and it includes definitions from other knowledge sources, such as WordNet [34]. This dataset provides comprehensive semantic information for the prompt templates we constructed, enabling the evaluation of both unstructured and structured sememe predictions simultaneously.

We selected 5000 English words from BabelSememe as our test set, and these words did not overlap with the dataset used in the experiments in Section 3.2.

Table 4. The first five entries of the BabelSememe dataset, which aligns the sememe definitions of senses in HowNet with words in WordNet and BabelNet, allowing us to obtain information from multiple knowledge sources simultaneously.

WordNet_ID	BabelNet_ID	HowNet_DEF
02417725a	00109443a	{watery 稀}
02417725a	02417725a	{sparse 疏}
00617748v	00087717v	{do 做:manner = {wrong 误}}
00617748v	00087717v	{err 出错}
01899360a	00104716a	{wise 智:adjunct = {neg 否}}

4.2. Experiment Settings

For the relevant parameters in sememe similarity calculation, we adhered to Li’s [6] experimental settings. The parameter μ was set to 1.6 because the root sememe contributes significantly to the similarity calculation. α and β were set to 0.4 and 0.6, respectively, representing the proportions of the Markov model and tree similarity in the calculation process. β was set to be larger because tree similarity considers the overall structure of the sememe tree. φ and δ were set to 0.7 and 0.3, respectively, representing the proportions of node similarity and edge similarity. φ was set to be larger because the semantic correlation between nodes and words is stronger than that of edges.

For setting the similarity threshold for word filtering, we needed to ensure an adequate number of candidate words in the semantic field as semantic references while also considering the interference of noisy words with the overall semantics. As shown in Table 5, when the similarity threshold is set to 0.2, the F1 score is the highest. Overall, the F1 score initially increases and then decreases with the increase in the similarity threshold. This is because a similarity threshold that is too low fails to effectively filter out noisy words, while an excessively high threshold may incorrectly filter out some words that are central to the semantic field.

Table 5. F1 scores under different similarity filtering thresholds.

Similarity Threshold	F1
0	64.38
0.1	65.26
0.2	67.70
0.3	67.41
0.4	66.33
0.5	63.34
0.6	62.59
0.7	62.34
0.8	61.10
0.9	59.57

The bold represents the optimal value

4.3. Baselines

We evaluated our method on two tasks: unstructured sememe prediction and structured sememe prediction.

For the first task, where we output the sememes of the central word sense as an unordered set, the selected baseline methods were as follows:

- SPWE [22]: A word embedding-based method that recommends sememes based on similar words in the word vector space.
- CSP [23]: An ensemble model consisting of four sub-models, including SPWCF and SPCSE using internal word information and SPWE and SPSE using external word information.
- LD + seq2seq [35]: A sequence-to-sequence model utilizing the text definition.
- ScorP [27]: A model that predicts sememes by considering the local correlation between the various sememes of a word and the semantics of the different words in the definition.
- ASPSW [25]: A model that improves sememe prediction by introducing synonyms and derives an attention-based strategy to dynamically balance knowledge from synonym sets and word embeddings.

For the second task, the selected baseline methods were:

- P-RNN [36]: A model that uses a recursive neural network (RNN) and a multi-layer perceptron (MLP) to gradually generate edges and nodes in a depth-first order for a given word pair with root sememes.
- NSTG [32]: A model that generates sememe trees by multiplying probabilities of sememe pairs from synonym sets, using sentence-BERT encoding for definition information, and computing the similarity levels.
- TasTG [32]: A model that converts sememe trees into sequences through depth-first traversal, introduces the tree attention mechanism, divides the attention computation into semantic attention and positional attention, encodes definitions using BERT-base, and generates sememe sequences using a transformer decoder.

4.4. Evaluation Metrics

For the first task, we used the F1 score as the evaluation metric, a commonly adopted method in previous studies. Given the sets of sememes S for the target word and $\hat{S}$ for the central word, we calculated the precision (P) and recall (R) as follows:

$$P = \frac{\text{size}(\tilde{S} \cap S)}{\text{size}(\tilde{S})}, \quad R = \frac{\text{size}(\tilde{S} \cap S)}{\text{size}(S)} \quad (9)$$

The F1 score was then computed using the following formula:

$$F1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

For the second task, to measure the structural similarity between sememe trees, we used three metrics proposed by Ye [32]: the strict F1, edge, and vertex. Additionally, since the root node plays a crucial role in sememe trees, we introduced a new metric, root, to assess whether the root nodes of two sememe trees are consistent. Its definition is as follows:

$$\text{Root}(t1, t2) = \begin{cases} 1, & t1_{\text{root}} = t2_{\text{root}} \\ 0, & t1_{\text{root}} \neq t2_{\text{root}} \end{cases} \quad (11)$$

4.5. Main Experimental Results

4.5.1. Unstructured Sememe Prediction

Table 6 presents the results of our proposed SFCW method and the other baseline models on the test set for unstructured sememe prediction. The sememe set predicted by SFCW achieves the highest F1 score of 67.7%, outperforming the second-ranking method, ASPSW, by 6.7% and surpassing the last-ranking method, LD + seq2seq, by 39.2%.

Table 6. Experimental results of the unstructured sememe prediction task.

Method	F1
SPWE+SPSE [22]	40.25
CSP [23]	46.53
LD + seq2seq [35]	25.85
SCorP [27]	56.32
ASPSW [25]	61.02
SFCW	67.70

The bold represents the SOTA result

4.5.2. Structured Sememe Prediction

Table 7 shows the results of our proposed SFCW method and the other baseline models on the test set for structured sememe prediction. Since P-RNN is given the root node before prediction, we did not compare its root metric.

Table 7. Experimental results of the structured sememe prediction task.

Method	Root	Strict	Edge	Vertex
P-RNN [36]	-	50.22	54.53	60.28
NSTG [32]	29.25	26.18	26.82	32.11
TaSTG [32]	37.14	39.75	41.22	48.25
SFCW	66.51	59.60	60.32	67.70

The bold represents the SOTA results

From the results in Table 7, we can see that SFCW achieves the highest scores on the root metric (66.5%), strict F1 metric (59.6%), edge metric (60.3%), and vertex metric (67.7%) among all the methods. The strict F1 scores for NSTG and TaSTG are 26.1% and 39.7%, respectively, showing a reduction of 33.5% and 19.9% compared to SFCW. The correctness of the root node plays a decisive role in the quality of the sememe tree. P-RNN benefits from a given root node, resulting in a strict F1 metric of 50.2%, an edge metric of 54.5%, and a vertex metric of 60.2%, ranking second only to SFCW.

5. Discussion

5.1. Impact of Selection Strategies

In Section 3.2, we conducted confirmatory experiments on the semantic field compatibility of target word senses for different types of semantic fields. Based on the results, we formulated a strategy for selecting semantic fields. To verify the effectiveness of our strategy, we compared it with combinations of several semantic fields.

The experimental results are shown in Figure 6. The F1 scores of our selection strategy are significantly higher than those of combinations involving different types of semantic

fields. This validates our viewpoint in Section 3.2 that different categories of semantic fields exhibit distinct semantic tendencies, and these differences in semantic tendencies may introduce interference when selecting the central word. This also indicates that an appropriate selection strategy can help us obtain more accurate central words by narrowing down the scope of calculations in different scenarios.

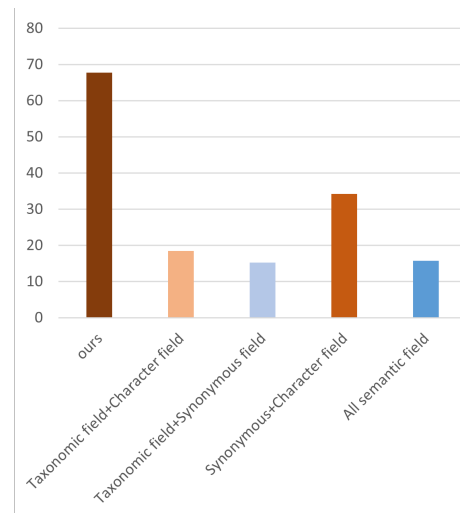


Figure 6. F1 scores under different semantic field selection strategies.

5.2. Central Word Sememe Structure

To better integrate the central word into sememe-related tasks, we conducted further analyses of its performance under different complexities.

We explored the optimal hierarchical structure of sememes by comparing the front n layers of the central word with the sememe tree of the target word. The experimental results, shown in Figure 7a, reveal a gradual decrease in the strict F1's structural similarity with increasing depth of the sememe tree. We observed that sememes affecting structural similarity are mainly concentrated in the second and third layers of the sememe tree, aligning with the primary depths in HowNet's sememe tree. Based on these findings, we recommend selecting sememes from the first two layers as the foundational sememe tree, allowing for better integration of the central word's sememes into other sememe-related tasks.

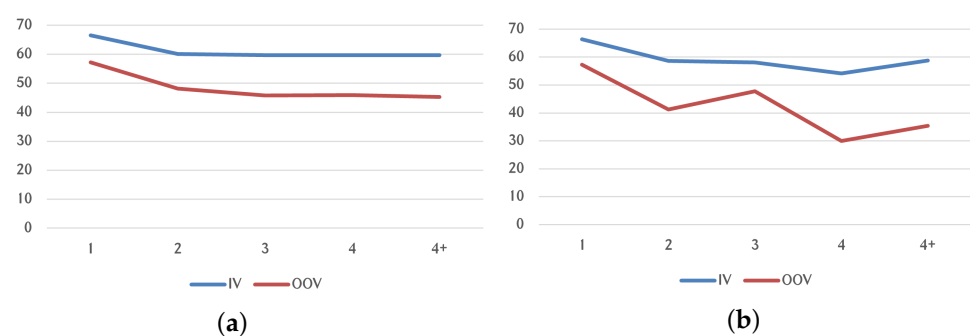


Figure 7. Test results on the complexity of trees. (a) Strict F1 of the first n levels of a sememe tree. (b) Strict F1 of a sememe tree of size n .

Furthermore, we investigated the impact of different sememe tree sizes on structural similarity, as depicted in Figure 7b. As the number of sememes in the sememe tree increases, the strict F1 score shows a fluctuating downward trend. However, it is worth noting that when the number of sememes is within 3, the strict-F1 value remains at a relatively high

level. Integrating the central word's sememes into other sememe-related tasks is thus more valuable within this range.

5.3. Advantages and Limitations of Semantic Fields

Compared to existing methods, SFCW has the following advantages:

(1) No training is required, and it exhibits a favorable time complexity ($O(n \times m)$) and space complexity ($O(n + m)$), where n is the number of predicted words, and m is the number of sememes per word. For a test set comprising 5000 words, the prediction task takes approximately 30 s.

(2) Previous methods primarily predicted sememes by computing word vector similarity or mapping word vectors. While vector-based methods can achieve performance improvements by incorporating the internal and external information of words, there always exists a semantic deviation between word vectors and sememe vectors. SFCW, by not explicitly mapping word vectors and sememe vectors but relying on semantic field matching between words, achieves better alignment between words and sememes.

(3) For structured sememe prediction, SFCW utilizes the sememe tree of the central word's sense as the prediction result, ensuring the integrity of the sememe structure.

However, using semantic fields as a medium for matching words and sememes has its limitations:

(1) The generation of semantic fields through ChatGPT leads to uncontrollable results.

(2) To broaden the applicability of semantic fields, we constructed various types of these fields. However, apart from taxonomic fields, the prediction results for other semantic fields are unstable.

(3) In structured sememe prediction, the accuracy of sub-nodes cannot be guaranteed.

5.4. Case Study

5.4.1. Unstructured Sememe Cases

Figure 8 illustrates the sememe set of “water meter” and its synonyms, hypernyms, and split characters. In the case of “water meter”, its near words in the taxonomic field, synonymous field, and character field exhibit some similarity in their sememe sets because there exists a certain level of semantic correlation between these semantic fields and the target words.

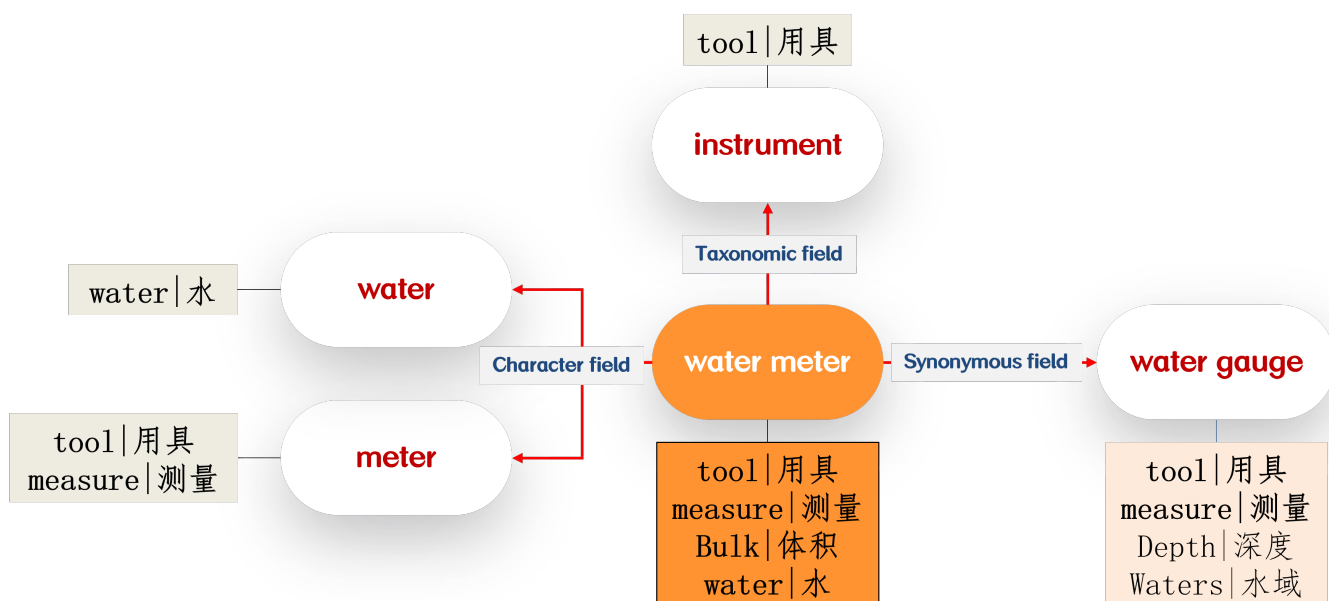


Figure 8. Sememe sets of the word “water meter” and its candidates.

The synonym “water gauge” and the character “meter” demonstrate the highest similarity, followed by “instrument” and “water”. This aligns closely with the prioritization of our semantic field selection strategy.

It is worth noting that all candidates, except for “water”, share the same element, “tool | 用具”. This is because the central word in “water meter” is “meter”, which conveys the key semantic meaning, while “water” serves a modifying role.

5.4.2. Structured Sememe Cases

Tables 8 and 9 include the sememe tree and average similarity of each sense for the candidates for “water meter”. The senses of different words and different senses of the same word in the table exhibit notable differences. However, if there exists a sense that shares certain semantics with the entire set of candidates, it becomes possible to find highly similar senses from the remaining set of candidates to match it. This allows the sense to achieve a high average sememe similarity.

Table 8. Average similarity of the candidate senses (taxonomic field and synonymous field).

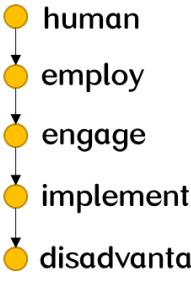
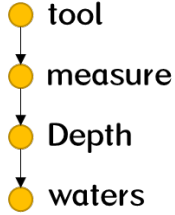
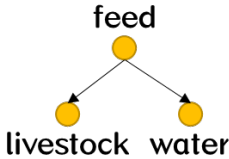
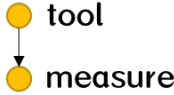
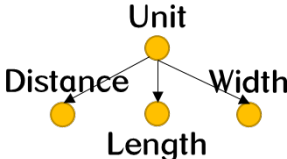
Type	Candidate Words	Candidate Senses	Sememe Tree	Avg Sim
Taxonomic field	instrument	乐器 (instrument): any of various devices or contrivances that can be used to produce musical tones or sounds	 MusicTool	0.56
		仪器 (instrument): a device that requires skill for proper use	 tool	0.67
		工具 (instrument): the mean whereby some act is accomplished	 implement	0.55
		爪牙 (instrument): a person used by another to gain an end	 human employ engage implement disadvantageous	0.39
Synonymous field	water gauge	水尺 (water gauge): gauge for indicating the level of water in e.g. a tank or boiler or reservoir	 tool measure Depth waters	0.70

Table 9. Average similarity of the candidate senses (character field).

Type	Candidate Words	Candidate Senses	Sememe Tree	Avg Sim
Character field	water	水 (water): binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid		0.45
		浇水 (water): supply with water, as with channels or ditches		0.26
		饮 (water): provide with water		0.24
	meter	仪表 (meter): any of various instruments for measuring a quantity		0.70
		米 (meter): the basic unit of length adopted under the System International Unites		0.16
		韵律 (meter): rhythm as given by division into parts of equal duration		0.16

For instance, consider the synonym sense “水尺 (water gauge)”, which shares sememes with the split character sense “仪表 (meter)” and the hypernym sense “仪器 (instrument)”. These senses demonstrate a significant level of sememe similarity (0.70, 0.67, and 0.70).

On the other hand, senses such as “米 (meter)” and “爪牙 (instrument)” have a low average similarity (0.16 and 0.39) since they do not share sememes with the candidates. As for the sense “水 (water)”, although it shares a common sememe element “water” with the target word “water meter”, their semantic tendencies do not align completely. This is because “water” serves as the root node in the sense “水 (water)”, whereas it functions as a modifier rather than the root node in the sense “水表 (water meter)”. As a result, the sememe expresses a central semantic meaning for the former sense but acts as a modifier for the latter, leading to a lower average similarity (0.45) in the table.

6. Conclusions

In this paper, we propose a sememe prediction method based on the central word of a semantic field. We use ChatGPT to acquire words with specific semantic relationships to construct the semantic field and devise a strategy for selecting semantic fields. Finally, by leveraging the semantic associations between senses and computing the average sememe similarity, we predict the central word’s sememes and utilize them as the prediction result. Compared to the existing methods, our approach achieves a superior semantic alignment

between words and sememes without the need for training. Moreover, it can be applied seamlessly to both unstructured and structured sememe prediction tasks.

We achieve optimal results in both unstructured and structured sememe prediction tasks. We also conduct qualitative and quantitative analyses of the sememe tree structure of the central word, illustrating how to integrate the central word more effectively into other HowNet-related tasks.

In the future, we plan to explore the integration of retrieval-augmented generation (RAG) techniques to address the issue of unpredictability in ChatGPT. Simultaneously, we aim to construct more stable semantic fields to further enhance the accuracy of predictions.

Author Contributions: Conceptualization, G.L. and Y.C.; methodology, G.L.; software, G.L.; validation, G.L.; formal analysis, G.L.; investigation, G.L.; resources, Y.C.; data curation, G.L.; writing—original draft preparation, G.L.; writing—review and editing, Y.C.; visualization, G.L.; supervision, Y.C.; project administration, Y.C.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSTL (grant number: 2023XM42).

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: [https://github.com/thunlp/BabelNet-Sememe-Prediction/blob/master/BabelSememe/synset_sememes.txt] (accessed on 24 March 2023)].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bloomfield, L. A Set of Postulates for the Science of Language. *Language* **1926**, *2*, 153–164. [[CrossRef](#)]
2. Dong, Z.; Dong, Q. HowNet—A hybrid language and knowledge resource. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 21–23 August 2003; pp. 820–824. [[CrossRef](#)]
3. Niu, Y.; Xie, R.; Liu, Z.; Sun, M. Improved Word Representation Learning with Sememes. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, CC, Canada, 30 July–4 August 2017; Barzilay, R., Kan, M.Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 2049–2058. [[CrossRef](#)]
4. Fan, M.; Zhang, Y.; Li, J. Word similarity computation based on HowNet. In Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China, 15–17 August 2015; pp. 1487–1492. [[CrossRef](#)]
5. Hu, F.S.; Guo, Y. An improved algorithm of word similarity computation based on HowNet. In Proceedings of the 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), Zhangjiajie, China, 25–27 May 2012; Volume 3, pp. 372–376. [[CrossRef](#)]
6. Li, H.; Zhou, C.L.; Jiang, M.; Cai, K. A hybrid approach for Chinese word similarity computing based on HowNet. In Proceedings of the International Conference on Automatic Control and Artificial Intelligence (ACAI 2012), Xiamen, China, 3–5 March 2012; pp. 80–83. [[CrossRef](#)]
7. Duan, X.; Zhao, J.; Xu, B. Word Sense Disambiguation through Sememe Labeling. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, 6–12 January 2007; IJCAI’07; pp. 1594–1599.
8. Hou, B.; Qi, F.; Zang, Y.; Zhang, X.; Liu, Z.; Sun, M. Try to Substitute: An Unsupervised Chinese Word Sense Disambiguation Method Based on HowNet. In Proceedings of the 28th International Conference on Computational Linguistics, Online, 8–13 December 2020; Scott, D., Bel, N., Zong, C., Eds.; International Committee on Computational Linguistics: Barcelona, Spain, 2020; pp. 1752–1757. [[CrossRef](#)]
9. Huang, M.; Ye, B.; Wang, Y.; Chen, H.; Cheng, J.; Zhu, X. New Word Detection for Sentiment Analysis. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 22–27 June 2014; Toutanova, K., Wu, H., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 531–541. [[CrossRef](#)]
10. Wen, Z.; Gui, L.; Wang, Q.; Guo, M.; Yu, X.; Du, J.; Xu, R. Sememe knowledge and auxiliary information enhanced approach for sarcasm detection. *Inf. Process. Manag.* **2022**, *59*, 102883. [[CrossRef](#)]
11. Thakur, N. MonkeyPox2022Tweets: A Large-Scale Twitter Dataset on the 2022 Monkeypox Outbreak, Findings from Analysis of Tweets, and Open Research Questions. *Infect. Dis. Rep.* **2022**, *14*, 855–883. [[CrossRef](#)] [[PubMed](#)]
12. Negrón, J.B. #EULAR2018: the annual European congress of rheumatology—a twitter hashtag analysis. *Rheumatol. Int.* **2019**, *39*, 893–899. [[PubMed](#)]
13. Li, F.L.; Chen, H.; Xu, G.; Qiu, T.; Ji, F.; Zhang, J.; Chen, H. AliMeKG: Domain Knowledge Graph Construction and Application in E-Commerce. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM’20), New York, NY, USA, 19–23 October 2020; pp. 2581–2588. [[CrossRef](#)]
14. Qi, F.; Yao, Y.; Xu, S.; Liu, Z.; Sun, M. Turn the Combination Lock: Learnable Textual Backdoor Attacks via Word Substitution. 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), Online, 1–6 August 2021; Zong, C.; Xia, F.; Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4873–4883. [\[CrossRef\]](#)
15. Qiang, J.; Lu, X.; Li, Y.; Yuan, Y.; Wu, X. Chinese Lexical Simplification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1819–1828. [\[CrossRef\]](#)
 16. Gu, Y.; Yan, J.; Zhu, H.; Liu, Z.; Xie, R.; Sun, M.; Lin, F.; Lin, L. Language Modeling with Sparse Product of Sememe Experts. *arXiv* **2018**, arXiv:1810.12387.
 17. Qi, F.; Huang, J.; Yang, C.; Liu, Z.; Chen, X.; Liu, Q.; Sun, M. Modeling Semantic Compositionality with Sememe Knowledge. *arXiv* **2019**, arXiv:1907.04744.
 18. Qi, F.; Chang, L.; Sun, M.; Ouyang, S.; Liu, Z. Towards Building a Multilingual Sememe Knowledge Base: Predicting Sememes for BabelNet Synsets. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8624–8631. [\[CrossRef\]](#)
 19. Yang, L.; Kong, C.; Chen, Y.; Liu, Y.; Fan, Q.; Yang, E. Incorporating Sememes into Chinese Definition Modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1669–1677. [\[CrossRef\]](#)
 20. Qian, C.; Feng, F.; Wen, L.; Chua, T.S. Conceptualized and Contextualized Gaussian Embedding. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 13683–13691. [\[CrossRef\]](#)
 21. Qin, Y.; Qi, F.; Ouyang, S.; Liu, Z.; Yang, C.; Wang, Y.; Liu, Q.; Sun, M. Improving Sequence Modeling Ability of Recurrent Neural Networks via Sememes. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2364–2373. [\[CrossRef\]](#)
 22. Xie, R.; Yuan, X.; Liu, Z.; Sun, M. Lexical Sememe Prediction via Word Embeddings and Matrix Factorization. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 19–25 August 2017; pp. 4200–4206. [\[CrossRef\]](#)
 23. Jin, H.; Zhu, H.; Liu, Z.; Xie, R.; Sun, M.; Lin, F.; Lin, L. Incorporating Chinese Characters of Words for Lexical Sememe Prediction. *arXiv* **2018**, arXiv:1806.06349.
 24. Lyu, B.; Chen, L.; Yu, K. Glyph Enhanced Chinese Character Pre-Training for Lexical Sememe Prediction. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; Moens, M.F., Huang, X., Specia, L., Yih, S.W.t., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4549–4555. [\[CrossRef\]](#)
 25. Kang, X.; Li, B.; Yao, H.; Liang, Q.; Li, S.; Gong, J.; Li, X. Incorporating Synonym for Lexical Sememe Prediction: An Attention-Based Model. *Appl. Sci.* **2020**, *10*, 5996. [\[CrossRef\]](#)
 26. Bai, M.; Lv, P.; Long, X. Lexical Sememe Prediction with RNN and Modern Chinese Dictionary. In Proceedings of the 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, 28–30 July 2018; pp. 825–830. [\[CrossRef\]](#)
 27. Du, J.; Qi, F.; Sun, M.; Liu, Z. Lexical Sememe Prediction using Dictionary Definitions by Capturing Local Semantic Correspondence. *arXiv* **2019**, arXiv:1907.04744.
 28. Qi, F.; Lin, Y.; Sun, M.; Zhu, H.; Xie, R.; Liu, Z. Cross-lingual Lexical Sememe Prediction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 358–368. [\[CrossRef\]](#)
 29. WANG, H.; LIU, S.; DUAN, J.; HE, L.; LI, X. Chinese Lexical Sememe Prediction Using CilinE Knowledge. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2023**, *E106.A*, 146–153. [\[CrossRef\]](#)
 30. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [\[CrossRef\]](#)
 31. Liu, Y.; Qi, F.; Liu, Z.; Sun, M. Research on Consistency Check of Sememe Annotations in HowNet. *J. Chin. Inf. Process.* **2021**, *35*, 23–34. [\[CrossRef\]](#)
 32. Ye, Y.; Qi, F.; Liu, Z.; Sun, M. Going “Deeper”: Structured Sememe Prediction via Transformer with Tree Attention. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 128–138. [\[CrossRef\]](#)
 33. Qi, F.; Yang, C.; Liu, Z.; Dong, Q.; Sun, M.; Dong, Z. OpenHowNet: An Open Sememe-based Lexical Knowledge Base. *arXiv* **2019**, arXiv:1901.09957.
 34. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [\[CrossRef\]](#)
 35. Li, W.; Ren, X.; Dai, D.; Wu, Y.; Wang, H.; Sun, X. Sememe Prediction: Learning Semantic Knowledge from Unstructured Textual Wiki Descriptions. *arXiv* **2018**, arXiv:1808.05437.
 36. Liu, B.; Shang, X.; Liu, L.; Tan, Y.; Hou, L.; Li, J. Sememe Tree Prediction for English-Chinese Word Pairs. In Proceedings of the Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence, Guangzhou, China, 4–7 November 2021; Chen, H., Liu, K., Sun, Y., Wang, S., Hou, L., Eds.; Springer: Singapore, 2021; pp. 15–27.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.