

Article

Bilingual Extraction and Alignment of Indigenous Chinese Linguistic Terminology via Multi-Channel Graph Neural Networks

Hongyue Diao ^{1,2}, Zongyu Zhang ¹, Sihan Ji ¹ and Hao Wei ^{1,2,*} ¹ School of Software, Dalian University of Foreign Languages, Dalian 116044, China; diaohongyue@dlufl.edu.cn (H.D.)² China Research Center for Northeast Asian Languages, Dalian University of Foreign Languages, Dalian 116044, China

* Correspondence: weihao@dlufl.edu.cn

Abstract

Terms are specialized words and expressions used in particular disciplines, cultures, or fields. They usually carry precise meanings and aim to describe referents accurately and clearly. Due to differences in culture, history, and other factors across countries, the development of indigenous Chinese linguistic terms plays a vital role in bridging cultural gaps and promoting the dissemination of Chinese culture. These terms not only explain specific words in Chinese and describe unique linguistic phenomena, but also embody the core concepts and academic traditions of Chinese linguistics, thereby contributing to the global spread and development of Chinese civilization. In order to achieve cross-linguistic dissemination of indigenous terms, we construct a linguistically informed bilingual corpus encompassing a broad spectrum of linguistic subfields, together with novel methods for the automatic extraction and cross-linguistic alignment of terminologies. The resulting corpus contains over 22,000 aligned sentence pairs across nine linguistic domains, providing a robust foundation for bilingual term mining. Building upon this resource, we further propose a multi-channel graph neural network (MCGNN) that jointly models semantic, syntactic, sequential, and co-occurrence relations, thereby enabling multi-perspective reasoning and achieving more accurate bilingual term extraction and alignment. Experimental results demonstrate that our approach substantially improves the accuracy and consistency of bilingual term extraction, alleviates the resource scarcity in the linguistic domain, and provides a solid foundation for future research and applications in cross-linguistic knowledge sharing and academic communication.

Keywords: indigenous Chinese linguistic terminology; term extraction; term alignment; deep learning; graph neural network

Academic Editor: Douglas
O'Shaughnessy

Received: 3 February 2026

Revised: 27 February 2026

Accepted: 27 February 2026

Published: 3 March 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As global exchanges between civilizations become increasingly frequent, language serves as the core vehicle for cultural communication. In this context, the accurate transmission and comprehension of indigenous linguistic terms have become ever more crucial. Terms are essential linguistic units for building professional knowledge systems, acting as key vehicles for expressing domain-specific concepts and ideas. Given the importance of terminology in various disciplines such as translation studies, information retrieval, and semantic modeling, term extraction is gradually becoming one of the core tasks in both

linguistic and interdisciplinary research in computational linguistics. Traditional terminology research has largely relied on linguists manually annotating corpora and building structured databases. Although this approach ensures high quality, it encounters significant challenges in efficiency and consistency when applied to cross-lingual, large-scale, and multi-domain scenarios. Term extraction [1] refers to the process of identifying and retrieving indigenous terms from Chinese-language literature using techniques such as natural language processing. Its core resides in the precise demarcation and identification of specialized terminologies. Term alignment [2] signifies the process of systematically correlating indigenous Chinese terms with their counterparts in other languages. It entails transcending cultural disparities and achieving accurate semantic mapping through semantic association modeling. In this paper, we construct a linguistically informed bilingual corpus of indigenous terms to support automatic term extraction and alignment across languages, covering a wide range of linguistic subfields. Additionally, we propose a novel multi-channel graph neural network (MCGNN) that integrates four complementary perspectives—co-occurrence patterns, syntactic structures, semantic relations, and sequential dependencies—via dedicated graph encoders (GCN, GAT, GraphSAGE, and a BiLSTM with relative position embeddings). This design enables the model to capture both local and global contextual signals, thereby enhancing the robustness and accuracy of term identification and alignment. To the best of our knowledge, this is the first work that jointly models these four linguistic views within a unified graph-based framework for bilingual terminology processing, going beyond simple multi-view concatenation by using specialized aggregation mechanisms tailored to each relation type. The entire model is trained end-to-end with a CRF decoder, treating term extraction and alignment as a unified sequence labeling problem. The joint modeling of term extraction and bilingual alignment contributes to model performance in several ways. First, unlike pipelined systems that perform extraction and alignment in separate stages, our unified sequence labeling approach encodes each Chinese–English sentence pair as a single contiguous sequence with a separator token, allowing the graph neural network to propagate information across language boundaries during message passing. This cross-lingual information flow means that extraction in one language can benefit from the alignment signal in the other: when the model identifies a term on the English side, the shared representation space and graph structure help constrain and refine term boundaries on the Chinese side, and vice versa. Second, alignment quality depends critically on correct extraction, since aligned term types are derived from the predicted BIO labels on both sides. By jointly optimizing both objectives in one forward pass, the model learns to balance boundary precision and cross-lingual consistency, rather than suffering from error propagation that plagues pipelined approaches. Third, the experimental results reflect this synergy: MCGNN achieves the highest aligned accuracy and align F1 among all baselines, while also attaining the best monolingual term extraction F1-scores for both Chinese and English. In contrast, models that perform well on extraction alone do not necessarily excel on alignment metrics, indicating that the integrated design captures cross-lingual correspondences that isolated extraction or post hoc alignment cannot. Moreover, we also introduce a new term-level alignment strategy to complement the category-level alignment scheme in conventional sequence labeling-based alignment tasks, providing a more fine-grained evaluation. These findings demonstrate the potential of combining linguistically informed resources with advanced neural architectures for scalable and effective terminology processing.

2. Related Work

2.1. Monolingual Term Extraction

Automatic term extraction (ATE) refers to the task of automatically identifying domain-specific terms from large-scale text corpora, and it has long been regarded as a fundamental problem in ontology construction, text summarization, and knowledge graph development [3]. Previous studies have also highlighted its importance in downstream applications such as information retrieval and machine translation [4]. According to the techniques and principles of monolingual term extraction, existing approaches can generally be categorized into five types: rule-based methods, statistical methods, machine learning-based methods, traditional deep learning-based methods, and, more recently, large language model (LLM)-based methods [3,4]. Each of these approaches has its own advantages, and integrating multiple methods has been shown to further improve overall performance [5].

2.2. Bilingual Term Extraction and Alignment

According to the type of corpora used, bilingual term extraction can generally be divided into two main approaches: methods based on parallel corpora and methods based on comparable corpora. Parallel corpora consist of pairs of source- and target-language texts that are mutual translations, exhibiting a high degree of alignment. This property allows parallel corpora to provide direct and reliable contrastive information, making them particularly suitable for identifying bilingual term correspondences [6]. In contrast, comparable corpora are not mutual translations but consist of monolingual texts in different languages covering the same or similar topics. Compared with parallel corpora, comparable corpora are easier to obtain and offer broader coverage, but they pose greater challenges for alignment [7].

Bilingual term extraction has evolved through various methodological perspectives, depending on the availability and quality of linguistic resources. While some studies exploit comparable corpora to broaden domain coverage, others rely on parallel corpora to achieve more precise alignment. The choice of corpus type often reflects a trade-off between data accessibility and alignment accuracy. In this work, we focus on bilingual linguistic parallel corpora, as they offer stronger contextual correspondence and higher reliability for identifying translation equivalents.

Lefever et al. [8] present a language-pair-independent approach that uses sub-sentential alignment and statistical filtering to extract high-quality term pairs across multiple languages, demonstrating strong performance for cross-language information processing. Vintar [9] proposes a Slovene–English term mapping method combining statistical and syntactic extraction from word-aligned parallel texts, noting that the syntactic method captures high-value multi-word terms but requires labeled corpora. Zhang et al. [10] introduce a multi-level termhood measure derived from domain-general corpus comparisons to constrain bilingual alignment, effectively improving extraction accuracy. Almeida et al. [11] demonstrate the feasibility of building large-scale bilingual terminology databases by first constructing parallel corpora from websites, offering a practical solution for resource-poor languages. Ideue et al. [12] compare three unsupervised statistical measures (Score_F, Score_L, Score_C) for ranking candidate term pairs, showing that their combination significantly enhances the identification of correct pairs. Macken et al. [13] present the TExSIS system, which uses chunk-based alignment with statistical filtering and highlights the value of multilingual evidence in improving term unithood reliability. Guo et al. [14] train a Siamese network with hard negatives to map sentences into a shared embedding space, achieving competitive parallel corpus reconstruction at both sentence and document levels. Repar et al. [15] integrate embedding, character, and cognate features for high-precision term alignment, retrieving top candidates via cosine similarity before clas-

sification. Liu et al. [16] introduce GIST, a large-scale multilingual terminology dataset using a hybrid LLM–human framework, which informs our corpus construction and evaluation framework by validating the effectiveness of hybrid methodologies for terminology collection. Zaratiana et al. [17] propose GraphER, which formulates entity and relation extraction as graph structure learning, enabling dynamic refinement of graph structure during extraction and joint prediction rather than separate pipelines—an approach that resonates with our multi-channel graph design. Hu and Xu [18] present DM-BLI, applying dynamic multi-subspace alignment to bilingual lexicon induction, which aligns with our alignment supervision strategy by highlighting the importance of fine-grained, subspace-level cross-lingual correspondence. Latouche et al. [19] propose BinaryAlign, modeling word alignment as binary sequence labeling, which parallels our task formalization and evaluation approach, though our framework extends this to multi-category terminology with composite label structures.

Despite notable advancements in prior work on term extraction and alignment, existing methods still face several limitations that our work aims to address: (1) Existing methods often depend on manually constructed lexicons, heuristic rules, or pre-trained models. As a result, they tend to perform poorly in emerging domains or when encountering novel terms, revealing limited adaptability. (2) Most existing approaches struggle to adequately capture terms within their contextual relationships across semantic, syntactic, and document-level layers, particularly leading to errors in multi-word terms or semantically ambiguous cases. (3) In terms of bilingual term alignment, existing methods often rely on word- or phrase-level correspondences and neglect higher-level semantic consistency, which may result in inaccurate alignment of term pairs. To overcome these issues, we propose a novel neural network approach that integrates a multi-channel graph neural network framework from four complementary perspectives: co-occurrence patterns, syntactic structures, semantic relations, and sequential dependencies. This design enables our model to capture both local and global contextual signals, thereby enhancing the robustness and accuracy of term extraction and alignment in diverse domains.

In summary, the main contributions of this work are as follows:

1. We presented a hierarchical indigenous terminology system of Chinese linguistics and constructed the corresponding bilingual parallel corpus of indigenous terms. This taxonomy provides a systematic framework for organizing linguistic knowledge across multiple subfields, including nine major categories such as phonetics, morphology, semantics, and pragmatics. The resulting bilingual corpus, annotated under this unified framework, serves as a valuable resource for research in linguistic knowledge representation, terminology extraction, and cross-lingual language processing.
2. We proposed a multi-channel graph neural network (MCGNN) that jointly addresses term extraction and alignment by treating them as a unified sequence labeling problem. This is achieved through the integration of four distinct channels: syntax, semantics, sequence, and co-occurrence graphs, each employing a dedicated graph encoder (GAT, GraphSAGE, BiLSTM with relative position embeddings, and GCN). This integration allows the model to capture complementary linguistic and structural information, leading to more accurate term extraction and alignment.
3. We validate the effectiveness of the proposed MCGNN framework and the utility of the newly constructed bilingual linguistic corpus. Through systematic comparisons against a wide range of baselines, we demonstrate that MCGNN achieves consistent and statistically significant improvements in both term extraction and alignment tasks. Ablation studies, significance testing, and error analysis further corroborate the robustness of our model. In addition, we introduce a term-level evaluation paradigm

based on a composite labeling strategy, providing a new benchmark for future research in fine-grained terminology alignment.

3. Bilingual Corpus for Chinese Linguistic Indigenous Terminology

3.1. Construction of Terminological Type System

In this paper, to ensure comprehensive coverage and terminological accuracy, we systematically reviewed a large number of representative works in the field of Chinese linguistics, including textbooks, monographs, and journal articles. Based on this extensive review, we established a hierarchical terminology system comprising nine fundamental categories. This classification not only incorporates commonly used categories in traditional linguistics, but also covers representative categories of mainstream modern linguistics. Such a design ensures that the system inherits classical theoretical foundations while accommodating the applicability and coverage of modern linguistic research. These nine categories are essential for organizing linguistic knowledge and serve as the backbone for constructing domain-specific corpora, as well as facilitating terminology extraction, alignment, and knowledge mapping in the field.

Table 1 presents the complete bilingual classification system for Chinese linguistic terminology, listing the nine categories both in Chinese and their corresponding English equivalents. This structured classification provides a clear framework to standardize terminology usage and supports further research and application in Chinese linguistics.

To provide a clearer view of the linguistic terminology system, we also present a hierarchical graph illustrating the organization and relationships among the main categories and their representative example. The graph is shown in Figure 1.

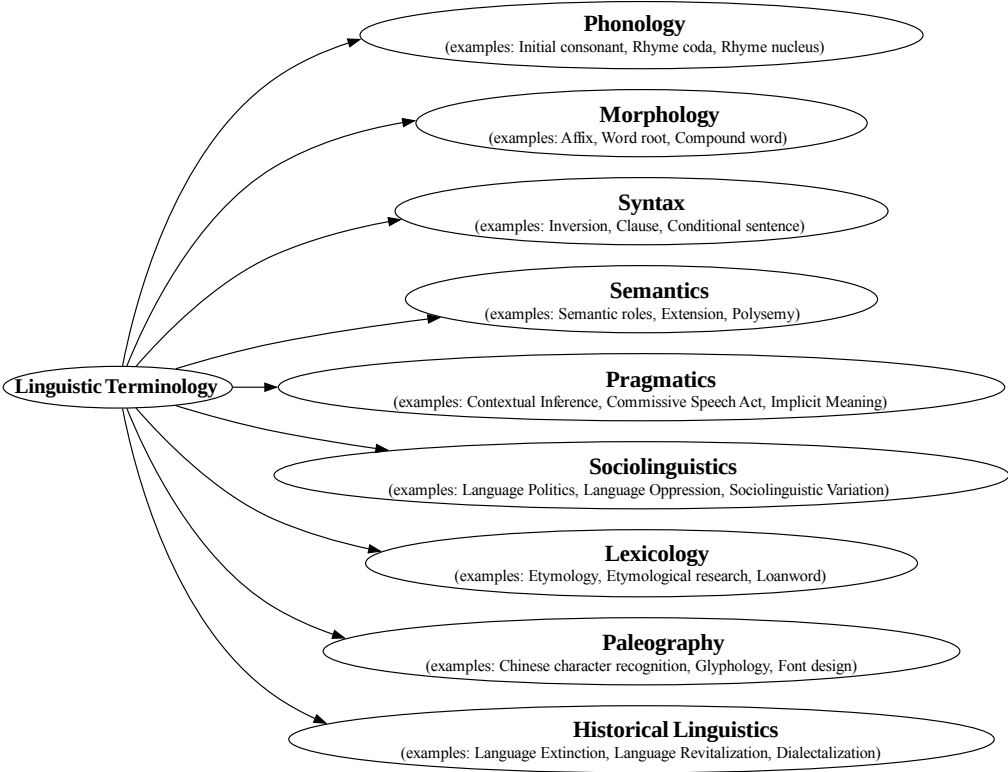


Figure 1. Hierarchical Structure of Linguistic Terminology.

Table 1. Term Classification System for Chinese Linguistic Terminology (Bilingual Version).

Category (CN)	Category (EN)
音系学	Phonology
文字学	Paleography
词汇学	Lexicology
句法学	Syntax
形态学	Morphology
语义学	Semantics
语用学	Pragmatics
社会语言学	Sociolinguistics
历史语言学	Historical Linguistics

3.2. Terminology Labeling

In this paper, term extraction and alignment are essentially a sequence labeling task, which requires proper annotation of raw data to enable effective model training. We adopt the widely used BIO (Begin–Inside–Outside) labeling scheme, where “B” marks the beginning of a term, “I” indicates the continuation inside the term, and “O” denotes tokens outside any term. Each sentence pair, consisting of Chinese text followed by its English translation, is separated by a <sep> marker, and each word/token is annotated with its corresponding BIO label, indicating different terminology types. In addition, we have annotated “align_id” to represent the fine-grained alignment link relationship between Chinese and English terms. The same Chinese and English IDs in a sentence represent a set of term pairs.

Building upon the nine major categories defined in the Chinese Linguistic Classification System introduced earlier, we generate a total of 19 distinct label classes to cover the different types of terms. When performing term-level alignment tasks, merge BIO labels with alignment IDs to generate composite labels. The maximum number of fine-grained term alignment type labels has been expanded from 19 to $18 \times n + 1$, where “n” is the maximum value of the same term type label in a sentence of the bilingual corpus. To illustrate this annotation approach, Table 2 presents an example of a labeled sequence from “Syntax” text samples. For instance, the sentence “Understanding the structure of subject is the basis for reading and writing” is annotated such that each token receives a corresponding BIO label reflecting its role and category in the term sequence. This example demonstrates how terms belonging to specific classes are precisely marked, facilitating accurate term extraction and alignment in subsequent processing.

Table 2. Example of annotating Bilingual Linguistics Parallel Corpus in Chinese and English.

Sequence	Label	Align_id	Sequence	Label	Align_id
理	O		Understanding	O	
解	O		the	O	
主	B-Syn	1	structure	O	
语	I-Syn	1	of	O	
的	O		subject	B-Syn	1
结	O		is	O	
构	O		the	O	
是	O		basis	O	
阅	O		for	O	
读	O		reading	O	

Table 2. *Cont.*

Sequence	Label	Align_id	Sequence	Label	Align_id
和	O		and	O	
写	O		writing	O	
作	O				
的	O				
基	O				
础	O				

3.3. The Static Analysis of Bilingual Corpus

Through a combination of automated annotation and manual verification, a bilingual Chinese linguistic corpus of indigenous terms was constructed. The detailed distribution of terms across different linguistic categories is presented in Table 3. Overall, the corpus contains 22,828 sentence pairs with a total of 136,872 annotated terms. Among these, phonetic terms (Pho) and morphological terms (Mor) account for the largest proportions, reflecting the richness of phonetic and morphological variation in the dataset. The number of “B” and “I” labels varies across categories, indicating differences in term length and complexity among the various linguistic types.

Table 3. Statistics of Terms and Labels by Category.

Category	Number of Samples	Total Terms	B Labels	I Labels
His	455	1372	509	863
Lex	370	1134	544	590
Mor	5852	36,434	14,181	22,253
Pal	253	1109	497	612
Pho	9935	64,361	25,226	39,135
Pra	720	3047	1117	1930
Sem	678	1743	715	1028
Soc	2178	8024	3010	5014
Syn	1667	8639	3695	4944
Total	22,828	136,872	56,294	80,527

4. Multi-Channel Graph Neural Network

4.1. Model Framework

Based on the architecture of Multi-Channel Graph Neural Networks, our framework for linguistic term extraction integrates four types of graphs—semantic graph, syntactic graph, sequence graph, and co-occurrence graph—as illustrated in Figure 2. The entire model consists of three main components: the embedding layer, the graph fusion layer, and the CRF layer. In the embedding layer, we integrate position embeddings and fused semantic embeddings, which are obtained by concatenating static word embeddings with contextualized dynamic embeddings, to comprehensively capture sequential dependencies, contextual relationships, and lexical features of the input text. The graph fusion layer aggregates multi-view information from the four heterogeneous graphs, each representing the input from a distinct linguistic perspective. Specifically, the semantic graph captures global contextual relationships derived from deep semantic similarity. The sequence graph encodes local word dependencies using BiLSTM and relative positional encoding to preserve sequential order. The syntactic graph incorporates grammatical structure through dependency parsing, while the co-occurrence graph reflects statistical co-appearance patterns among terms. By jointly modeling these complementary sources of linguistic knowledge,

the graph fusion layer generates enriched, structure-aware node representations. These representations are then passed to a CRF decoding layer, which performs structured prediction to identify and label term boundaries with high precision.

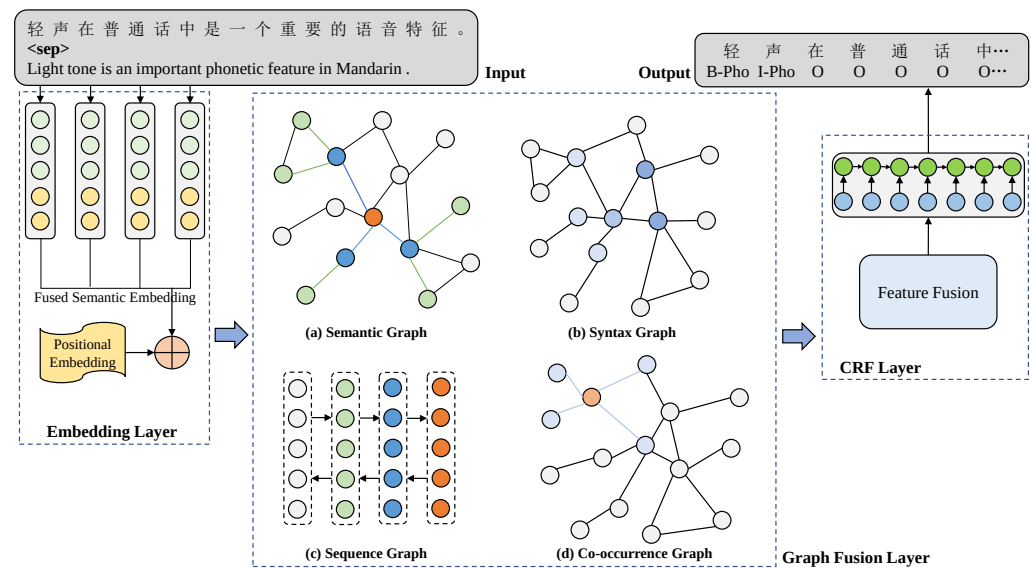


Figure 2. Model Architecture of Our Proposed Framework.

4.2. The Embedding Layer

4.2.1. Positional Embedding

Context information is a crucial factor influencing the model's performance. We introduce a positional embedding mechanism to encode the relative or absolute position of tokens within a sentence, enabling the model to better capture local and global dependencies. Specifically, for an input sequence of length n , the positional embedding matrix $P \in \mathbb{R}^{n \times d}$ is added to the input word embeddings $E \in \mathbb{R}^{n \times d}$:

$$X = E + P \quad (1)$$

Here, d denotes the embedding dimension. Each position i in P is assigned a fixed vector, which can be either learned or constructed using deterministic functions such as sine and cosine:

$$P_{(i,2k)} = \sin\left(\frac{i}{10,000^{2k/d}}\right), \quad P_{(i,2k+1)} = \cos\left(\frac{i}{10,000^{2k/d}}\right) \quad (2)$$

This design ensures that the model is aware of the order of the tokens, which is essential for downstream tasks such as linguistic terminology extraction. In our graph convolutional setup, these position-aware features are further propagated through the graph to enhance the representation of syntactic and semantic relations.

4.2.2. Fused Semantic Embedding

The input to our model consists of bilingual Chinese–English sentence pairs, where each word/token is annotated with its corresponding BIO label, indicating terminology types. Each sentence pair is concatenated using a <sep> marker, with Chinese text appearing first, followed by its English translation. This sequential format ensures that the model can simultaneously leverage both local context within each language and cross-lingual correspondences between languages.

Traditional static word embeddings are often insufficient for bilingual term extraction due to their inability to capture contextual variations and cross-lingual semantic alignment. To address this, we propose a multi-vector fusion approach that integrates static word embeddings, contextualized embeddings from XLM-RoBERTa-base, and positional encodings. By concatenating these three complementary representations, the model can leverage both the semantic generality of static embeddings and the contextual sensitivity of dynamic embeddings, while positional information further enhances structural awareness. This combined representation improves semantic alignment between Chinese and English terms, leading to more accurate and robust domain-specific terminology extraction, especially in low-resource or noisy environments.

4.3. The Graph Fusion Layer

4.3.1. Semantic Extraction Graph

The semantic graph is constructed based on the semantic similarity between words, where edges are formed by computing cosine similarities between word embeddings. To model the semantic relationships effectively, we adopt the GraphSAGE framework, which is well-suited for large-scale graph-structured data. Our semantic graph incorporates two key components: a Mean Aggregation layer and a Transformer layer. The Mean Aggregation layer aggregates information by computing the average features of a node's neighbors, which ensures high computational efficiency and memory friendliness. This enables scalable and efficient learning over massive semantic graphs. To enhancing the ability of expressing node, model also assembling Tranformer layer, optimizing node features by attention mechanism. The whole network uses multi-layer architecture, including batch normalization and dropout regularization, to guarantee the stability of training and enhance the model's generalization capability. The advantages of semantic graph is having ability to capture deep semantic relationships across words, providing semantic-level support for term recognition, maintaining high computing efficiency in the meantime.

$$\tilde{v}_i = \frac{v_i}{\|v_i\|_2} \quad (3)$$

The semantic graph framework begins by constructing a graph where each node represents a word with embedding vector v_i .

$$S_{i,j} = \tilde{v}_i \cdot \tilde{v}_j \quad (4)$$

Edges are established based on the cosine similarity $S_{i,j}$ between normalized embeddings $\tilde{v}_i$, and the adjacency matrix $A_{i,j}$ is determined by a similarity threshold θ .

$$A_{i,j} = \begin{cases} 1, & S_{i,j} > \theta \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

To efficiently capture and propagate semantic information, the model employs a multi-layer architecture that integrates both a Mean Aggregation layer and a Transformer layer.

$$h_i^{(k)} = \sigma \left(W^{(k)} \cdot \text{CONCAT} \left(h_i^{(k-1)}, \frac{1}{|N(i)|} \sum_{j \in N(i)} h_j^{(k-1)} \right) \right) \quad (6)$$

In the Mean Aggregation layer, each node's representation $h_i^{(k)}$ is updated by aggregating the average features of its neighbors $\mathcal{N}(i)$, ensuring scalability and computational efficiency.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (7)$$

The subsequent Transformer layer leverages the self-attention mechanism $\text{Attention}(Q, K, V)$ to further enhance node features, enabling the model to capture complex and long-range semantic dependencies.

Additional techniques such as batch normalization and dropout are utilized throughout the network to stabilize training and improve generalization. This design allows the semantic graph to effectively support term recognition by modeling deep semantic relationships while maintaining high performance and efficiency. The threshold θ for the semantic similarity graph was set to 0.6, which provided a good trade-off between graph sparsity and connectivity.

4.3.2. Syntax Extraction Graph

The syntax graph is constructed based on dependency syntax analysis, where syntactic dependency relations between words are extracted using the spaCy library for both Chinese and English texts. This graph captures deep structural relationships at the syntactic level and provides structural support for term recognition. The syntax graph leverages a Graph Attention Network (GAT) for feature learning, where an attention mechanism is employed to adaptively learn the importance of different syntactic dependencies. Each GAT layer contains learnable attention weights that dynamically adjust the strength of information propagation based on the type and intensity of the syntactic relations. In addition, the model incorporates dependency relation type encoding, which maps various syntactic relations (e.g., subject–verb, verb–object) into numerical representations. This provides the graph neural network with richer structural information. Through multi-head attention, the model is capable of simultaneously focusing on multiple syntactic patterns. The strength of the syntax graph lies in its ability to capture syntactic dependencies between words, making it particularly effective in identifying domain-specific terms with complex grammatical structures. It offers fine-grained syntactic guidance for accurate term recognition.

Figure 3 illustrates the structure of the Graph Attention Network (GAT) applied to the Syntax Extraction Graph.

$$A_{i,j} = \begin{cases} 1, & \text{if word } i \text{ and } j \text{ have a dependency relation} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

To construct the syntax graph, the adjacency matrix $A_{i,j}$ is determined by the existence of a syntactic dependency relation between word i and word j .

$$T_{i,j} = \text{dep_type_to_id}(\text{dep}(i,j)) \quad (9)$$

Each edge in the graph is assigned a dependency relation type encoding $T_{i,j}$, where the dependency type between word i and j is converted into a numerical identifier using a mapping function $\text{dep_type_to_id}(\cdot)$.

$$e_{ij} = \text{LeakyReLU}\left(a^\top [Wh_i \parallel Wh_j]\right) \quad (10)$$

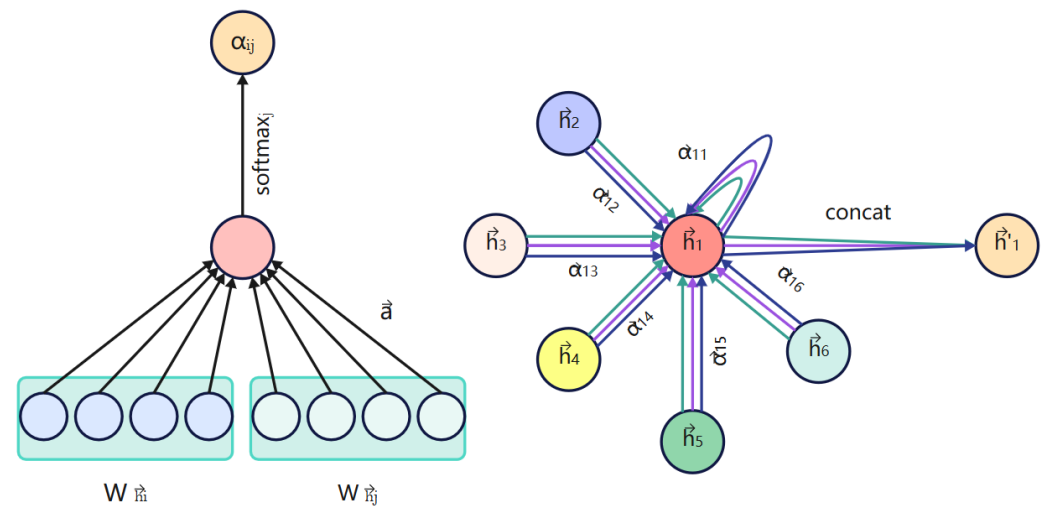


Figure 3. Graph Attention Network.

Within each GAT layer, the unnormalized attention coefficient e_{ij} between node i and its neighbor j is computed through a shared linear transformation followed by a LeakyReLU activation.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})} \quad (11)$$

The attention weights α_{ij} are obtained by normalizing e_{ij} across all neighbors of node i using a softmax function, ensuring the relative importance of each neighbor is preserved.

$$h'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j \right) \quad (12)$$

Each node's feature representation h'_i is updated by aggregating its neighbors' transformed features, weighted by the learned attention coefficients α_{ij} , followed by a non-linear activation σ .

$$h'_i = \parallel_{m=1}^M \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^m W^m h_j \right) \quad (13)$$

To enhance learning capacity, multi-head attention is applied. Representations from multiple independent attention heads (M in total) are concatenated to form the final output, allowing the model to capture diverse syntactic patterns.

4.3.3. Sequence Extraction Graph

To capture sequential dependencies and fine-grained positional relationships among words, we construct a local sequence graph based on relative positions. While traditional sequence models like BiLSTM are not graph-based in nature, we integrate graph-inspired structural design by explicitly modeling relative positional relations between words in the sequence. Specifically, we treat each word as a node and connect it to other words within a fixed-size contextual window, forming a local neighborhood akin to a graph structure. The key innovation lies in the use of learnable relative position embeddings to represent edge relations between word nodes.

The overall architecture integrates BiLSTM for contextual encoding and a relative position module for position-aware representation. The BiLSTM provides rich semantic features, while the relative position embeddings introduce structural bias that mimics graph

edge weights. Together, they form a hybrid structure that combines sequence modeling with local graph connectivity.

$$H = \text{BiLSTM}(X) \quad (14)$$

The input word embeddings X are first passed through a BiLSTM to obtain context-aware representations H . This captures sequential semantics in both forward and backward directions.

$$R_{i,j} = \text{RelPosEmb}(i - j) \quad (15)$$

To encode the relative structural relations between word i and word j , we introduce a learnable embedding function $\text{RelPosEmb}(i - j)$, which generates a position embedding matrix $R_{i,j}$ based on the distance $i - j$.

$$r_i = \frac{1}{n} \sum_{j=1}^n R_{i,j} \quad (16)$$

For each word node i , its final position representation r_i is obtained by averaging its relative embeddings with all other nodes. This implicitly encodes the word's position in the local sequence graph.

$$O_i = \text{Dropout}(H_i + r_i) \quad (17)$$

Finally, the semantic and structural representations are fused via element-wise addition, followed by dropout. The output O_i captures both sequential semantics and local graph-inspired positional information.

4.3.4. Co-Occurrence Extraction Graph

The co-occurrence graph is constructed based on the co-occurrence relations of words within the text, aiming to capture local contextual dependencies between words. A sliding window mechanism is applied to count co-occurrence frequencies and build the co-occurrence matrix. Specifically, given a predefined window size (set to 5 by default), the graph computes how frequently each word co-occurs with surrounding words within the window, which is used to determine the edge weights between word nodes. The graph adopts a Graph Convolutional Network (GCN) for feature learning, where multiple layers of graph convolutions are employed to aggregate information from neighboring nodes. The architecture consists of an input layer and multiple hidden layers, each incorporating graph convolution, linear transformation, non-linear activation, and dropout regularization. To enhance the model's effectiveness, a TF-IDF weighting scheme is also introduced to reweight the co-occurrence matrix, emphasizing the contribution of more informative word pairs. The key advantage of the co-occurrence graph lies in its ability to effectively capture local semantic associations between words, thereby providing rich contextual information for term recognition.

$$C_{i,j} = \sum_{s \in S} \sum_{w_k=i} \sum_{\substack{w_l=j \\ |k-l| \leq w, k \neq l}} 1 \quad (18)$$

The co-occurrence matrix $C_{i,j}$ is computed by sliding a fixed-size window (width w) over all sentences $s \in S$ and counting how often word i co-occurs with word j within the window. This forms the basis of edge weights in the graph.

$$\text{TFIDF}(C) = \text{TfidfTransformer}(C) \quad (19)$$

To emphasize the importance of informative word pairs, the co-occurrence matrix C is transformed using a TF-IDF weighting scheme, producing a reweighted matrix that reduces the impact of frequent but uninformative co-occurrences.

$$A_{i,j} = \begin{cases} C_{i,j}, & C_{i,j} > \theta \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

Edges in the graph are determined by applying a threshold θ to the TF-IDF weighted co-occurrence matrix. Only pairs with values exceeding the threshold are retained in the adjacency matrix A .

$$\hat{A} = D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}} \quad (21)$$

To stabilize training and ensure numerical consistency, the adjacency matrix A is symmetrically normalized with degree matrix D , and self-loops are added via the identity matrix I .

$$H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}) \quad (22)$$

Feature propagation in the GCN is performed by repeatedly applying graph convolution layers. At each layer l , the hidden features $H^{(l)}$ are updated via a linear transformation $W^{(l)}$, followed by aggregation using the normalized adjacency matrix $\hat{A}$ and a non-linear activation function σ . For the co-occurrence graph, we applied TF-IDF weighting and retained edges with weight > 0.3 (after normalization) to filter out noisy co-occurrences.

4.4. The CRF Layer

To enhance the accuracy of sequence labeling and ensure global consistency, we incorporate a Conditional Random Field (CRF) layer on top of the fused feature representations. The CRF models the conditional probability of the entire output label sequence given the input features, capturing the dependencies between adjacent tags and effectively reducing invalid tag transitions.

Let $X = (x_1, x_2, \dots, x_n)$ denote the input sequence and $Y = (y_1, y_2, \dots, y_n)$ the corresponding predicted label sequence. The CRF defines the score of the sequence Y as

$$s(X, Y) = \sum_{i=1}^n (P_{i,y_i} + T_{y_{i-1},y_i}) \quad (23)$$

where P_{i,y_i} is the emission score of assigning label y_i to position i , produced by the previous encoder layers, and T_{y_{i-1},y_i} is the transition score from label y_{i-1} to y_i , representing the learnable transition probability matrix.

The CRF computes the conditional probability of a label sequence Y given the input X as

$$P(Y | X) = \frac{\exp(s(X, Y))}{\sum_{Y' \in \mathcal{Y}(X)} \exp(s(X, Y'))} \quad (24)$$

where $\mathcal{Y}(X)$ denotes the set of all possible label sequences for X . During training, the objective is to maximize the log-likelihood of the correct label sequence. In inference, the optimal label sequence $\hat{Y}$ is obtained by maximizing the score:

$$\hat{Y} = \arg \max_{Y' \in \mathcal{Y}(X)} s(X, Y') \quad (25)$$

This decoding is efficiently performed using the Viterbi algorithm.

By modeling both local (emission) and global (transition) interactions, the CRF layer effectively addresses label inconsistency issues, such as illegal transitions (e.g., ‘I-term’

without a preceding ‘B-term’), thereby improving the structural integrity and overall precision of term boundary predictions.

5. Experiment

5.1. Hardware Configuration

To evaluate the predictive performance of the proposed model, both our approach and several baseline methods were implemented and tested on a Windows 10 platform using Python 3.10. The model was developed using the PyTorch 2.1 framework, with training conducted on an NVIDIA RTX 4090 GPU under CUDA 11.1. All experiments were run with a batch size of 64 and an initial learning rate of 2×10^{-5} , optimized using the Adam optimizer. For reproducibility, random seeds were set for Python, NumPy, and PyTorch. The GPU utilization was monitored to ensure efficient memory usage and computation during training. In addition, data preprocessing and augmentation procedures were applied consistently across all methods to ensure fair comparison. The experimental setup allows replication of the results and provides a benchmark for evaluating the effectiveness of the proposed model against baseline approaches, as shown in Table 4.

Table 4. Experimental Platform Configuration.

Component	Specification
Operating System	Windows 10
Programming Language	Python
Deep Learning Framework	PyTorch
CUDA Version	11.1
GPU Model	NVIDIA RTX 4090

5.2. Parameter Settings

In this paper, we employ a multi-graph fusion approach for bilingual terminology recognition. The model integrates four different types of graph neural networks: co-occurrence graph (CoGCN) for local word patterns, semantic graph (GraphSAGE) for semantic relationships, syntax graph (GAT) for grammatical dependencies, and sequence graph (BiLSTM) for sequential information. For dependency analysis, we employ spaCy’s pre-trained models: zh_core_web_sm for Chinese and en_core_web_sm for English, which provide accurate syntactic dependency trees for both languages. The entire corpus of 22,828 sentence pairs was randomly divided into training (80%, 18,262 pairs) and test (20%, 4566 pairs) sets. The split was stratified by term category to ensure similar distributions across all subsets. To avoid data leakage, the semantic graph and co-occurrence graph are constructed exclusively on the training set. For test sentences, we treat each sentence as an isolated graph: nodes are the tokens of that sentence, and edges are determined solely by within-sentence relations (co-occurrence within a window) or by connecting to training-set nodes only if the token appears in the training vocabulary (using pre-computed similarities). The syntactic graph and sequence graph are built on-the-fly for each sentence independently, using the dependency parser and relative position encoding, so no leakage occurs. The specific details are shown in Table 5.

Table 5. Parameter Settings.

Component	Specification
<i>Training Parameters</i>	
Optimizer	Adam
Initial Learning rate	2×10^{-5}
Epochs	50
Batch size	64
Gradient accumulation steps	2
<i>Model Architecture</i>	
Pre-trained model	XLM-RoBERTa-base
Embedding dimension	768
Hidden dimension	256
Dropout rate	0.1
Maximum sequence length	256
<i>Label Configuration</i>	
Number of labels	19
<i>Co-occurrence Graph</i>	
Window size	5
Number of layers	2
<i>Semantic Graph</i>	
Aggregator class	MeanAggregator
Neighbor sampling numbers	10
<i>Sequence Graph</i>	
Max relative position	8
<i>Syntax Graph</i>	
Dependency parser (Chinese)	zh_core_web_sm
Dependency parser (English)	en_core_web_sm

5.3. Comparison Model

To further demonstrate the advantages of our model, we conduct a comprehensive comparison with traditional neural networks and large language models (LLMs). In addition, we introduce two new model combinations—XLM-RoBERTa + BiLSTM + CRF (XRBC) and XLM-RoBERTa + Transformer + CRF (XRTC)—which use the same XLM-RoBERTa embedding followed by BiLSTM or Transformer encoder and CRF layer, but without any graph module. These two new combination baselines directly measure the contribution of multi-graph fusion. The detailed comparison is presented as follows:

1. RNN [20]: RNN is a classical sequence modeling method for time-series data, using an updated hidden state to remember past info and handle variable-length sequences. However, it is hard to parallelize and prone to vanishing or exploding gradients.
2. LSTM [21]: LSTM is an enhanced RNN for capturing long-term dependencies in sequential data, using memory cells and three gates to selectively manage information and avoid vanishing gradients. However, its complex structure causes higher computational cost and slower training than simpler models like GRU.
3. BiLSTM [22]: BiLSTM captures context bidirectionally for complex semantic dependencies but has high computational cost and limited parallelism, especially with long sequences.
4. BiGRU [23]: BiGRU is a simplified LSTM variant using gates for information flow, boosting efficiency while keeping core sequence modeling abilities—though it is slightly less expressive than LSTM.

5. TextGRU [24]: TextGRU is a lightweight GRU-based recurrent model for text representation, which handles variable-length sequences by controlling information flow via update and reset gates. Compared to LSTM, it is more computationally efficient while keeping most sequence modeling capabilities, and fits tasks like sentiment analysis and text classification.
6. BERT [25]: BERT is a Transformer-based pre-trained model that captures deep bidirectional context, excelling at NLP tasks via fine-tuning.
7. ERNIE [26]: ERNIE extends BERT by integrating structured knowledge from knowledge graphs/entities, enhancing semantic capture and performance on tasks needing world knowledge or entity understanding.
8. RoBERTa [27]: RoBERTa is a BERT improvement that optimizes pre-training (more data, longer training, no next-sentence prediction, dynamic masking) for stronger, more robust representations in downstream NLP tasks.

5.4. Evaluation Metrics

An essential step to evaluate performance on different models is to select appropriate metrics. We employ two primary types of evaluation. First, for the standard sequence labeling task within each language, we adopt the widely used metrics of Precision (P), Recall (R), and F1-score. These are calculated based on a strict matching criterion, where a predicted entity is considered a true positive (TP) only if both its boundaries and its type exactly match a ground-truth entity. Precision and Recall are formally defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (26)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (27)$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The F1-score, representing the harmonic mean of Precision and Recall, is defined as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (28)$$

Second, to specifically measure the model's capability in identifying corresponding terms across bilingual sentence pairs, we introduce the Term Alignment metrics (which measure category-level and term-level consistency) for comparison. For each sentence pair s , let $T_{\text{align_true}}(s)$ be the set of ground-truth-aligned term types (i.e., types that appear in both the Chinese and English sentences), and $T_{\text{align_pred}}(s)$ be the set of predicted aligned types. The term-level alignment metrics use the same calculation method as the category-level, but the labels are matched based on composite labels (a combination of BIO labels and align_id). We define the following:

- Term Alignment Accuracy ($\text{Acc}_{\text{align}}$): The proportion of sentence pairs where the predicted aligned type set exactly matches the ground-truth set, among all pairs with at least one aligned type:

$$\text{Acc}_{\text{align}} = \frac{\sum_{s \in S_{\text{aligned}}} \mathbb{I}(T_{\text{align_pred}}(s) = T_{\text{align_true}}(s))}{|S_{\text{aligned}}|} \quad (29)$$

where S_{aligned} is the set of sentence pairs with non-empty $T_{\text{align_true}}(s)$, and $\mathbb{I}(\cdot)$ is the indicator function.

- Term Alignment Precision, Recall, and F1: These metrics are computed by micro-averaging over all sentence pairs with at least one aligned type:

$$\text{Precision}_{\text{align}} = \frac{\sum_{s \in S_{\text{aligned}}} |T_{\text{align_pred}}(s) \cap T_{\text{align_true}}(s)|}{\sum_{s \in S_{\text{aligned}}} |T_{\text{align_pred}}(s)|} \quad (30)$$

$$\text{Recall}_{\text{align}} = \frac{\sum_{s \in S_{\text{aligned}}} |T_{\text{align_pred}}(s) \cap T_{\text{align_true}}(s)|}{\sum_{s \in S_{\text{aligned}}} |T_{\text{align_true}}(s)|} \quad (31)$$

$$\text{F1}_{\text{align}} = 2 \times \frac{\text{Precision}_{\text{align}} \times \text{Recall}_{\text{align}}}{\text{Precision}_{\text{align}} + \text{Recall}_{\text{align}}} \quad (32)$$

Together, these metrics provide a comprehensive view of both term extraction performance and the model's ability to learn cross-lingual semantic alignments.

6. Results and Discussion

6.1. Performance Comparison with Existing Models

The proposed term extraction model and evaluation metrics are introduced above. Each model underwent at least five training sessions, and the average results obtained during the training period were reported. Table 6 presents the results.

Table 6. Comparison of Extraction and Alignment Results on Bilingual Term Corpus.

Model	P (CN)	R (CN)	F1 (CN)	P (EN)	R (EN)	F1 (EN)	AliAcc	AliP	AliR	AliF1
RNN	67.24	44.82	53.79	59.33	40.96	48.46	72.93	98.33	88.45	93.13
LSTM	79.77	60.76	68.98	75.84	48.67	59.29	84.19	98.81	93.49	96.08
BiLSTM	73.78	55.76	63.52	58.37	46.91	52.02	82.73	98.66	92.81	95.64
BiGRU	74.95	54.06	62.81	80.96	53.69	64.56	79.55	98.60	91.84	95.10
TextGRU	72.41	71.09	71.74	68.40	69.59	68.99	82.91	95.88	95.07	95.47
BERT	91.78	88.79	90.26	77.77	84.30	80.90	90.02	97.83	98.09	97.96
ERNIE	93.00	89.62	91.28	82.34	82.26	82.30	90.72	98.27	97.83	98.05
RoBERTa	93.25	90.80	92.01	79.71	84.36	81.97	91.66	98.41	98.04	98.22
XRBC	91.80	89.20	90.49	80.84	82.22	81.52	90.10	98.15	97.40	97.77
XRTC	92.55	89.95	91.23	82.02	82.71	82.36	91.19	98.32	98.01	98.16
MCGNN*	93.29	91.34	92.31	81.89	82.92	82.41	91.52	98.26	98.11	98.19
Full Model	94.46	91.45	92.93	82.25	84.54	83.38	92.24	98.63	98.14	98.39

The comparison results demonstrate that our proposed model outperforms various baseline models (such as RNN, BiGRU, BiLSTM, BERT, etc.) in both Chinese- and English-term extraction tasks. Among traditional neural networks, TextGRU achieves the best performance, with F1 scores of 71.74% and 68.99% for Chinese and English, respectively, highlighting its strength in capturing contextual information and long-distance dependencies. For large-scale pre-trained models, BERT, ERNIE, and RoBERTa all surpass traditional neural networks. XLM-RoBERTa + BiLSTM + CRF and XLM-RoBERTa + Transformer + CRF have also achieved competitive results. In comparison, our Full Model achieves F1 scores of 92.93% for Chinese and 83.38% for English, corresponding to improvements of approximately 0.92% and 1.02% percentage points over the strongest baseline. Furthermore, it achieves the highest category-level alignment metrics (Aligned Accuracy 92.24%, Align F1 98.39%), underscoring the advantages of the integrated multi-channel graph neural network. Table 7 reports the introduced term-level alignment results for representative

baseline models. This experiment uses fine-grained composite labels with a larger number of categories and an imbalanced distribution. The full model achieves an F1 of 91.22%, which is lower than the category-level AliF1 (98.39%) but still significantly higher than the baselines, demonstrating its ability to capture fine-grained term correspondences. It is also evident that the model generally performs better on the Chinese corpus. Several factors may contribute to this observation. First, from a linguistic perspective, Chinese sentences tend to have more rigid word order and shorter average sentence lengths compared to English, facilitating the model's ability to capture contextual dependencies and term boundaries. Second, Chinese terminology often consists of fixed multi-character words or compounds, which are more consistent across contexts, whereas English terms exhibit greater morphological variation and flexible phrasal structures, increasing extraction difficulty. Third, the Chinese segmentation process is relatively standardized, reducing tokenization ambiguity, while English tokenization must handle spaces, punctuation, and compound words, potentially introducing errors. Finally, the training corpus may contain more balanced or abundant Chinese examples for certain term categories, providing stronger exposure for the model and contributing to higher F1 scores. Taken together, these linguistic and data-related factors explain why the model generally achieves better performance on the Chinese corpus than on the English one.

Table 7. Term-level Alignment Results (Selected Models).

Model	P _{term}	R _{term}	F1 _{term}
BERT	85.37	87.88	86.61
RoBERTa	90.98	88.93	89.95
XLM-BiLSTM-CRF	89.09	87.83	88.46
XLM-Transformer-CRF	89.47	90.23	89.86
Full Model	91.81	90.58	91.22

To further validate the generalization ability of our model, we evaluate it under a fully inductive setting where the semantic and co-occurrence graphs are constructed solely from each test sentence, without any connections to the training set. Specifically, for the semantic graph, we compute cosine similarities between the embeddings of tokens within the same sentence and connect pairs exceeding the same threshold; for the co-occurrence graph, we use the same window size to count intra-sentence co-occurrences, without any TF-IDF weighting derived from the training set. Baseline models are inherently inductive and their performance remains unchanged from Table 6; thus, we focus on the relative performance drop of our model. Under the inductive setting, MCGNN* achieves an F1 (CN) of 92.31%, an F1 (EN) of 82.41%, and an Align F1 of 98.19%. Compared to the original Full Model results (92.93%, 83.38%, and 98.39%, respectively), the relative performance drops are only 0.62, 0.97, and 0.20 percentage points. These marginal declines indicate that our model does not heavily rely on global corpus statistics; instead, the sentence-level syntactic, sequential, semantic, and co-occurrence information captured by the multi-channel graphs suffices to maintain high performance. Notably, while ablation experiments (Section 6.2) showed that removing the entire semantic graph causes a substantial drop, the much smaller declines under the fully inductive setting demonstrate that the sentence-level semantic relations captured by the intra-sentence graph already supply most of the semantic cues. Moreover, even in this purely inductive setting, our model maintains competitive performance, indicating that its multi-channel graph design effectively leverages local sentence structures to compensate for the absence of global priors.

Additionally, in terms of model training costs, compared to the baseline RoBERTa that achieved competitive results, MCGNN introduces four additional graph modules,

increasing total trainable parameters by about 10%. Training time per epoch grows from 5 min to 6.2 min on the existing hardware configuration, which is acceptable given the substantial performance gains.

6.2. Ablation Experiment

Given the complex structure of our model, it is essential to validate the effectiveness of each component. To this end, we performed ablation studies by removing each component individually and evaluating its contribution to the overall performance. The detailed results and metrics for each dataset are summarized in Table 8.

Table 8. Experimental Results of Ablation on Bilingual Term Corpus.

Model	P (CN)	R (CN)	F1 (CN)	P (EN)	R (EN)	F1 (EN)	AliAcc	AliP	AliR	AliF1
w/o SyG	95.06	87.08	91.91	81.32	83.11	82.21	90.58	98.57	97.41	97.99
w/o SemG	92.95	81.45	86.85	78.17	62.98	69.97	89.50	99.15	96.05	97.58
w/o SeqG	89.34	87.88	88.60	80.88	82.18	81.53	88.94	98.06	97.25	97.65
w/o CoG	96.86	89.07	92.83	81.72	83.71	82.71	91.66	98.45	97.90	98.17
Full Model	94.46	91.45	92.93	82.25	84.54	83.38	92.24	98.63	98.14	98.39

We present an ablation study evaluating the contribution of each component of our model on both Chinese and English datasets. Removing any component—Syntax Graph (SyG), Semantic Graph (SemG), Sequence Graph (SeqG), or Co-occurrence Graph (CoG)—leads to a decrease in performance across all metrics, demonstrating the importance of each module. Among these, removing the Semantic Graph causes the largest drop in F1 scores, from 92.93% to 86.85% for Chinese and from 83.38% to 69.97% for English, indicating its critical role in capturing hierarchical semantic relationships and domain-specific knowledge that are essential for accurate term extraction. The Syntax Graph contributes significantly by modeling structural dependencies between words; its removal reduces F1 to 91.91% for Chinese and 82.21% for English, showing that syntactic information helps the model delineate term boundaries and interpret complex sentence structures. The Co-occurrence Graph provides valuable statistical context on term co-occurrence patterns, leading to an F1 drop to 92.83% (Chinese) and 82.71% (English) when removed, which highlights its role in reinforcing frequent or related term relationships. The Sequence Graph, which encodes sequential dependencies, also plays an important role, with F1 decreasing to 88.60% for Chinese and 81.53% for English when omitted, suggesting that capturing local sequential patterns helps the model handle contiguous multi-word terms effectively. Overall, these results confirm that each graph module provides complementary information—semantic, syntactic, sequential, and co-occurrence-based—which collectively enhances both term extraction and alignment performance. The ablation study also explains why the model maintains robust performance in noisy or complex datasets: the combination of multiple perspectives allows it to compensate for missing information in any single view. Figure 4 provides a visual comparison of the performance drops under different ablation settings, clearly showing that the Semantic Graph contributes the most, followed by Co-occurrence, Syntax, and Sequence Graphs, respectively.

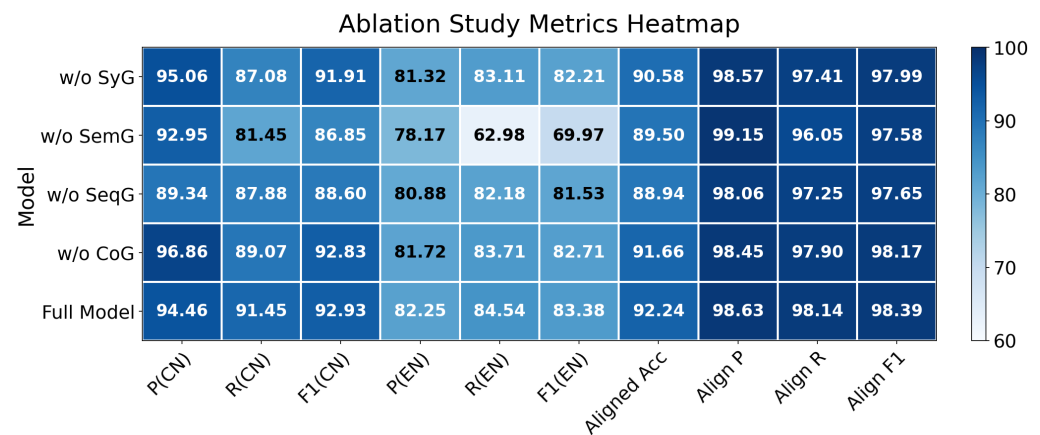


Figure 4. Visualization of performance comparison across different ablation settings.

6.3. Statistical Significance Test

To rigorously assess whether the performance gains of our full model are statistically significant, we conducted a paired bootstrap test with 10,000 resamples on the test set. For each resample, we computed the F1-scores of our model and a baseline on the same set of sentence pairs, and recorded the difference. As shown in Table 9, all improvements are significant at $p < 0.01$, confirming that the multi-channel graph fusion consistently yields better results.

Table 9. Statistical significance of performance differences between our full model and baselines.

Model	F1 (CN)	F1 (EN)	Significance
RNN	53.79	48.46	**
LSTM	68.98	59.29	**
BiLSTM	63.52	52.02	**
BiGRU	62.81	64.56	**
TextGRU	71.74	68.99	**
BERT	90.26	80.90	**
ERNIE	91.28	82.30	**
RoBERTa	92.01	81.97	**
XLM-BiLSTM-CRF	90.49	81.52	**
XLM-Transformer-CRF	91.23	82.36	**
Full Model	92.93	83.38	-

Note: ** indicates $p < 0.01$ (paired bootstrap test).

Among the pre-trained baselines, RoBERTa achieves the highest Chinese F1 (92.01%), while XLM-RoBERTa + Transformer + CRF attains the best English F1 (82.36%) among non-graph models. This reflects a known trade-off: single-language models like RoBERTa excel in their target language, whereas multilingual models like XLM-RoBERTa provide stronger cross-lingual transfer, benefiting the English side of our bilingual task. The Transformer-based encoder (XLM-RoBERTa + Transformer + CRF) outperforms its BiLSTM counterpart (XLM-RoBERTa + BiLSTM + CRF) by 0.74% (Chinese) and 0.84% (English), demonstrating the advantage of self-attention in modeling long-range dependencies crucial for term boundary detection.

Despite these strong baselines, our full model surpasses them by a clear margin. The bootstrap tests confirm that these differences are not due to chance. This underscores the value of integrating multiple linguistic perspectives (syntax, semantics, sequence, co-occurrence) through graph neural networks, which capture complementary information that even powerful Transformer-based models cannot fully exploit.

6.4. Error Analysis

In this section, we summarize and analyze the erroneous samples, selecting four main error types for quantitative and qualitative error analysis. First, we validate the effectiveness of our model by quantifying the proportions of error types in MCGNN and the RoBERTa baseline. Using the total number of errors made by RoBERTa across the four categories as the baseline (100%), Table 10 presents the error distribution on the test set. Thanks to multi-channel graph fusion, MCGNN reduces errors across all types, particularly boundary and type errors in term extraction. Cross-language misalignment is also reduced by 2.7 percentage points, indicating improved bilingual consistency.

Second, we conduct a qualitative analysis of several typical cases of prediction errors in MCGNN, thus further evaluating the directions for future improvement of our model. Upon examining the results, we find that the main sources of errors in bilingual terminology extraction arise from fundamental differences in how Chinese and English organize and express linguistic information. As shown in Table 11, several representative cases demonstrate how these cross-linguistic structural differences lead to boundary inconsistencies, category confusion, and partial alignment failures between the two languages.

Table 10. Quantitative Error Distribution (Percentage of Total Errors).

Error Type	Baseline	MCGNN
Boundary errors	38.2	31.3
Type errors	29.5	23.1
Omission errors	18.4	16.8
Cross-lingual misalignment	13.9	11.2

In the first example, “音位变体 / allophones,” the model successfully identifies the Chinese term but fails to accurately map it to the corresponding English counterpart. Although both represent the same phonological concept, the Chinese term is a compact compound noun, while the English equivalent is a single lexical unit embedded within a larger syntactic structure. This discrepancy highlights the model’s difficulty in establishing fine-grained semantic correspondence when the morphological density of Chinese contrasts with the lexical simplicity of English.

In the second case, “语用意义 / pragmatic meaning,” the model predicts the correct term span but assigns it to an incorrect category (“semantic” instead of “pragmatic”). Such misclassification occurs because both semantic and pragmatic terms often co-occur in overlapping contexts, making it challenging for the model to separate contextual meaning from pragmatic usage.

In the third example, “句法依存关系 / syntactic dependencies,” the model fails to correctly identify the core term on both sides. Although the phrase occurs in a clearly defined syntactic position, the internal dependency structure and surrounding modifiers in the sentence increase the model’s difficulty in boundary detection. This error reflects the model’s sensitivity to complex clause structures, where embedded modifiers and relational terms often obscure the precise span of linguistic concepts.

Finally, “文本处理模块 / text processing modules,” demonstrates a partial omission where the model captures the beginning of the compound but misses its full extent in both languages. This shows the model’s limitations in handling long compound terms that involve nested modification or technical expressions spanning several tokens.

Overall, these cases indicate that the main challenges in bilingual terminology extraction lie not in annotation inconsistencies, but in the intrinsic divergence of linguistic structures, lexical compounding, and semantic layering between Chinese and English. Future research will focus on designing more fine-grained cross-linguistic alignment mech-

anisms and enhancing the robustness of category-level discrimination to better capture such subtle but systematic discrepancies.

Table 11. Example of Error Analysis.

Ground-Truth (Annotated)	Model Prediction
[语音学] 研究 [音位模式]、[音素] 和 [音位变体] 在 [语流音变] 中的作用。 / [Phonology] analyzes [sound patterns], [phonemes], and [allophones] in [speech variation].	[语音学] 研究 [音位模式]、[音素] 和 音位变体 在 [语流音变] 中的作用。 / [Phonology] analyzes [sound patterns], [phonemes], and allophones in [speech variation].
[语义学] 探讨 [词义演变] 与 [语用意义] 在语境变化下的互动。 / [Semantics] investigates [meaning evolution] and [pragmatic meaning] under contextual variation.	[语义学] 探讨 [词义演变] 与 语用意义 在语境变化下的互动。 / [Semantics] investigates [meaning evolution] and pragmatic meaning under contextual variation.
[句法学] 分析 [句法依存关系] 在 [复合句结构] 中的可预测性。 / [Syntax] analyzes [syntactic dependencies] and their predictability in [complex sentence structures].	[句法学] 分析 句法依存关系 在 [复合句结构] 中的可预测性。 / [Syntax] analyzes syntactic dependencies and their predictability in [complex sentence structures].
[字形学] 的 计算机 应用 涵盖 从 [输入法设计] 到 [文本处理模块] 的多个层面。 / [Character morphology] applications in computing range from [input method design] to [text processing modules].	[字形学] 的 计算机 应用 涵盖 从 [输入法设计] 到 文本处理 模块 的多个层面。 / [Character morphology] applications in computing range from [input method design] to text processing modules.

7. Conclusions

To promote the dissemination of Chinese linguistic knowledge, this paper proposes a novel term extraction and alignment framework that integrates multiple graph-based models to comprehensively capture linguistic features from diverse perspectives, thereby enhancing the accuracy and robustness of indigenous Chinese linguistic terms identification. The graph fusion model effectively integrates syntactic, semantic, and lexical information through a multi-view graph structure, capturing complementary linguistic features across different levels of language representation. This multi-faceted approach enables the model to robustly identify terms by simultaneously considering local contextual cues, syntactic dependencies, semantic relations, and domain-specific knowledge, thereby significantly improving precision and recall in terminology extraction tasks. Furthermore, we constructed a comprehensive category system for indigenous Chinese linguistic terminology and developed a bilingual parallel corpus annotated under this taxonomy, providing valuable resources for future research in linguistic knowledge representation and computational linguistics. Experimental results validate the effectiveness of our proposed model and confirm its applicability to diverse linguistic scenarios. Overall, our work not only advances the development of term extraction and alignment techniques, but also contributes to the systematic construction of Chinese linguistic resources and the broader goal of automated language understanding.

Although we achieve promising results with the proposed term extraction framework on our linguistic dataset and validate the effectiveness of multi-graph fusion, there remain several directions worth exploring. First, MCGNN shows limited domain transferability, it is trained and evaluated solely on the Chinese linguistic terminology corpus, and its performance in cross-domain scenarios (e.g., medical, legal, or technical terminology) remains unverified. Second, the construction of multi-graph structures relies on sufficient linguistic resources and prior knowledge. In low-resource settings where parallel data

or high-quality parsers are scarce, the applicability of our approach may be constrained. Third, the error analysis reveals that the model occasionally fails to capture fine-grained or implicit terms, particularly those involving complex cross-lingual correspondences, morphologically dense Chinese compounds versus lexically simple English equivalents, category confusion between semantically proximate labels, and partial omission of long compound terms with nested modification.

In future work, we plan to enhance the framework by incorporating more fine-grained cross-linguistic feature alignment mechanisms, potentially leveraging bilingual word embeddings trained on large-scale parallel corpora to better capture semantic correspondences between Chinese and English terms. We also aim to design specialized modules to handle structural differences in term expression across languages, further improving the model's robustness and interpretability in real-world linguistic applications. Additionally, we will continue to refine and expand the taxonomy of indigenous Chinese linguistic terminology by incorporating more linguistic subfields and categories, thereby enhancing the comprehensiveness and systematicity of the classification framework.

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

Funding: This work was supported by the General Projects of the 14th Five-Year Plan of the National Language Commission (Grant No. YB145-82); Liaoning Provincial Department of Education project (Grant No. LJKQZ20222451 and JYTQN2023149); and the Natural Science Foundation of Liaoning Province of China (Grant No. 2024-BS-203).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Maynard, D.; Li, Y.; Peters, W. NLP Techniques for Term Extraction and Ontology Population. In *Ontology Learning and Population: Bridging the Gap Between Text and Knowledge*; IOS Press: Amsterdam, The Netherlands, 2008.
2. Ruder, S.; Vulić, I.; Søgaard, A. A survey of cross-lingual word embedding models. *J. Artif. Intell. Res.* **2019**, *65*, 569–631. [\[CrossRef\]](#)
3. Tran, H.T.H.; Martinc, M.; Caporusso, J.; Doucet, A.; Pollak, S. The recent advances in automatic term extraction: A survey. *arXiv* **2023**, arXiv:2301.06767. [\[CrossRef\]](#)
4. Xu, K.; Feng, Y.; Li, Q.; Dong, Z.; Wei, J. Survey on terminology extraction from texts. *J. Big Data* **2025**, *12*, 1–40. [\[CrossRef\]](#)
5. Chun, Y.; Kim, M.; Kim, D.; Park, C.; Lim, H. Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval. *arXiv* **2025**, arXiv:2506.21222. [\[CrossRef\]](#)
6. Zhang, M.; Wu, Q. Bilingual terminology extraction using multi-level termhood. *arXiv* **2013**, arXiv:1302.4492. [\[CrossRef\]](#)
7. Jia, H.; Li, C.; Wang, T. Bilingual terminology extraction from comparable e-commerce corpora. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021.
8. Lefever, E.; Macken, L.; Hoste, V. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 30 March–3 April 2009; pp. 496–504.
9. Vintar, S. Using parallel corpora for translation-oriented term extraction. *Babel* **2001**, *47*, 121–132. [\[CrossRef\]](#)
10. Zhang, C.; Wu, D. Bilingual terminology extraction using multi-level termhood. *Electron. Libr.* **2012**, *30*, 295–309. [\[CrossRef\]](#)
11. Almeida, J.J.; Simões, A. Automatic parallel corpora and bilingual terminology extraction from parallel websites. In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, Valletta, Malta, 22 May 2010.

12. Ideue, M.; Yamamoto, K.; Utiyama, M.; Sumita, E. A comparison of unsupervised bilingual term extraction methods using phrase-tables. In Proceedings of the Machine Translation Summit XIII: Papers, Xiamen, China, 19–23 September 2011.
13. Macken, L.; Lefever, E.; Hoste, V. Taxis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminol. Int. J. Theor. Appl. Issues Spec. Commun.* **2013**, *19*, 1–30. [\[CrossRef\]](#)
14. Guo, M.; Shen, Q.; Yang, Y.; Ge, H.; Cer, D.; Abrego, G.H.; Stevens, K.; Constant, N.; Sung, Y.H.; Strophe, B.; et al. Effective parallel corpus mining using bilingual sentence embeddings. *arXiv* **2018**, arXiv:1807.11906. [\[CrossRef\]](#)
15. Repar, A.; Martinc, M.; Ulcar, M.; Pollak, S. Word-embedding based bilingual terminology alignment. In Proceedings of the Electronic Lexicography in the 21st Century (eLex 2021) Post-Editing Lexicography, Virtual, 5–7 July 2021; p. 98.
16. Liu, J.; Ouzzani, I.; Li, W.; Zhang, L.; Ou, T.; Bouamor, H.; Jin, Z.; Diab, M.T. Toward Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset (GIST). In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025.
17. Zaradiana, U.; Tomeh, N.; El Khbir, N.; Holat, P.; Charnois, T. GraphER: A Structure-aware Text-to-Graph Model for Entity and Relation Extraction. *arXiv* **2024**, arXiv:2404.12491.
18. Hu, L.; Xu, Y. DM-BLI: Dynamic Multiple Subspaces Alignment for Unsupervised Bilingual Lexicon Induction. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024. [\[CrossRef\]](#)
19. Latouche, G.; Carbonneau, M.A.; Swanson, B. BinaryAlign: Word Alignment as Binary Sequence Labeling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024. [\[CrossRef\]](#)
20. Williams, R.J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1989**, *1*, 270–280. [\[CrossRef\]](#)
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
22. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [\[CrossRef\]](#)
23. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555. [\[CrossRef\]](#)
24. He, G.; Lu, W. Entire information attentive GRU for text representation. In Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, Tianjin, China, 14–17 September 2018.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019.
26. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129. [\[CrossRef\]](#)
27. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.

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.