International Journal for Housing Science and Its Applications
Publish August 3, 2025. Volume 46, Issue 3 Pages 1041-1054

https://doi.org/10.70517/ijhsa46380

A method for constructing and analyzing the knowledge graph of English learners in colleges and universities based on local linear embedding algorithm

Xiaoli Chen[1,*] and Ruijuan Hu[2]
[1] Zhengzhou Business University, Zhengzhou, Henan, 451200, China
[2] College English Department, Henan Finance of University, Zhengzhou, Henan, 450046, China
Corresponding authors: (e-mail: chenxl@zbu.edu.cn).

Abstract This study proposes a method for constructing and analyzing the knowledge graph of English learners in colleges and universities based on the local linear embedding LLE algorithm, which optimizes personalized learning support through dynamic characterization and cross-domain knowledge migration. The dynamic knowledge graph covering 9673 nodes is constructed with the core literacy of English discipline as the orientation, and the adjacency matrix (dimension 9673×9673) and 512-dimensional feature vector are generated using the feedback data of exercises. The joint knowledge migration method HLLEJKT is proposed to achieve cross-linguistic representation space alignment through kernelization extension and label optimization, and achieves 75.13% and 85.75% accuracy in XNLI and STS 2020 benchmark tasks, respectively, which is an improvement of 55.69 and 124 percentage points compared with the traditional method mBERT. In practical applications, the graph-based intelligent retrieval system achieves an average accuracy rate of 96.55% in the retrieval of teaching resources in English, Chinese, Spanish, and Arabic, with an accuracy rate of 98.47% in English and 94.56% in Arabic, and the length of the retrieval path is shortened to 21.60, which is 17.40% lower than that of the traditional method of fuzzy retrieval path for teaching resources. The method effectively integrates semantic association mining and knowledge migration mechanism, providing a theoretical breakthrough and practical paradigm for multilingual education technology.

Index Terms local linear embedding, LLE, knowledge graph, college English learning, knowledge migration

I. Introduction

In the wave of digital transformation, the rapid development of information technology is reshaping the field of education in an unprecedented way. With the advancement of the process of educational informatization, English teaching in colleges and universities has shifted from the traditional paper-based medium to a digital and intelligent form [1]. This shift not only greatly enriches the content and form of teaching resources, but also brings about the complexity of knowledge structure and information fragmentation [2]. Based on this background, in order to reduce the difficulty of teaching complex English knowledge, cultivate students' awareness of structured learning, and enable them to construct a more complete English knowledge system, the knowledge mapping tool in artificial intelligence technology can be utilized to carry out correlative English exploration activities.

Knowledge mapping is a combination of theories and methods in the fields of graphics, information science, co-occurrence analysis, etc., to present the knowledge structure graphically, intuitively, completely and logically [3]. In the process of teaching English in colleges and universities, teachers can make use of knowledge mapping to guide students in structured learning of English knowledge. Teachers let students read the textbook text combined with the knowledge map, understand the meaning of the text according to the expanded information to summarize the important and difficult knowledge in the text, so as to guide the students to carry out the correlation inquiry [4], [5]. Knowledge sorting and construction activities can also be carried out with the help of knowledge mapping, so that students can realize the systematic memory of English knowledge [6]. Let students use line segments, graphics, English, Chinese, symbols, etc., to build a net-like English knowledge structure, so that they can build a more logical and complete unit knowledge system by adding, subtracting and reorganizing the knowledge map given by the system [7], [8].

Knowledge mapping has been widely used in the process of education and teaching, and literature [9] proposes a knowledge mapping-driven method for student behavior prediction and multi-learning task recommendation, which fuses and associates the semantics of the extracted knowledge concepts with the multiple entities of students' learning behaviors in order to safeguard the accuracy of the prediction and recommendation tasks. Literature [10]

describes a chatbot for PBL (Project Based Learning) that integrates retrieval augmentation generation method and knowledge graph technology with personalized guidance, timely feedback, and accurate Q&A, which can provide a high learning experience for project-based teaching. It can be found that the application of knowledge mapping in the field of education is conducive to solving the current pain point problems in the retrieval of teaching resources, helping students to grasp the internal logic of knowledge as a whole, and improving learning efficiency.

Some scholars have also analyzed the impact of knowledge mapping on students' knowledge construction and the promotion of educational functions in English teaching. Literature [11] formulated a knowledge map featuring English teaching knowledge points, and combined it with the results of student behavioral data analysis collected by an online education platform to recommend English learning resources for students to meet their learning needs. Literature [12] applied a knowledge graph-based question and answer system to English teaching tests, providing accurate answers for English tests based on the flexible combinatorial structure and rich semantic expression ability of knowledge graphs, and avoiding the problem of missing answers that existed in previous question and answer systems. Literature [13] studied the construction and application of knowledge graph for English majors, and by extracting the relevant knowledge nodes of English major curriculum system from educational big data and constructing knowledge graph, it can effectively improve the existing teaching problems and optimize the quality of English teaching. Literature [14] studied the impact of knowledge graph-based course evaluation methods on college students' sense of acquisition of English learning, and knowledge graph-based learning context data can respond to students' learning evaluation from multiple dimensions and levels, and never improve their sense of acquisition in terms of knowledge, ability and emotion. For this reason, optimizing the knowledge graph construction and analysis process of English teaching by introducing intelligent algorithms is of great significance to improve the effect of English teaching.

This study proposes a dynamic and personalized construction method, combined with the local linear embedding LLE algorithm, to achieve the dual optimization of knowledge representation and migration. The initial map is constructed through the feedback data of exercises to capture the semantic association between knowledge points. And the local linear embedding technique is introduced to maintain the local structure of data in low-dimensional space, which enhances the characterization ability of the atlas to the individual learning state. On this basis, the joint knowledge migration method HLLEJKT is proposed to align the data distributions of the source and target domains through the mapping matrix, and optimize the intra- and inter-class distances by using the label information to achieve cross-language and cross-task knowledge migration. Same-class local linear embedding maintains the local linear structure of same-class samples in the low-dimensional space by reconstructing the weights through the nearest-neighbor nodes. The kernelization extension is extended to regenerate Hilbert space by introducing kernel functions to deal with nonlinear data, which enhances the model's ability to capture complex semantic relations. The method aligns the multilingual representation space with the English contrastive learning model through an alternate training strategy, which significantly improves the performance of cross-language tasks and effectively enhances the adaptability of the maps, providing a new idea for knowledge fusion in multilingual pre-trained models.

II. English Knowledge Graph Construction and Knowledge Migration Method Based on Local Linear Embedding

II. A. Methods of Knowledge Mapping Construction in English Subjects

As a reflection of the view of the essence of the discipline and the values of education in the discipline, the core qualities of the discipline point out the direction for the realization of the goal of nurturing people in the discipline, and at the same time provide specific guidance and support for the teaching of the discipline. We believe that in order to cope with the shortcomings in the development of subject knowledge mapping, the only way to truly bring the advantages of subject knowledge mapping into play is to construct knowledge maps oriented on subject core literacy, which can not only help students learn and understand subject knowledge systematically and promote the transfer and application of knowledge, but also provide powerful support for teachers in designing teaching and evaluating students' learning ability. It also provides strong support for teachers to design teaching and assess students' learning ability. The construction of subject mapping model under the development of disciplinary core literacy is to project the connotation of disciplinary core literacy and the educational and teaching needs for the development of disciplinary core literacy into specific subfields, and to integrate the teaching knowledge and teaching activities within the disciplinary field in order to adapt the knowledge mapping for the development of disciplinary core literacy. Based on this, this paper takes the English discipline as an example and proposes the construction method of the knowledge mapping of the English discipline under the guidance of the development of core literacy. We first elucidate the core literacy of the English discipline and clarify the construction goals; secondly, we analyze the structure of the English discipline and decompose the goals into different categories of learning

materials; finally, we combine the knowledge mapping technology with the principles of teaching and learning, and define the entity types, entity attributes, and entity relationships of the knowledge mapping of the English discipline in order to ensure the scientificity and accuracy of the model. The process of constructing the knowledge mapping of English subject under the orientation of core literacy development is shown in Figure 1.

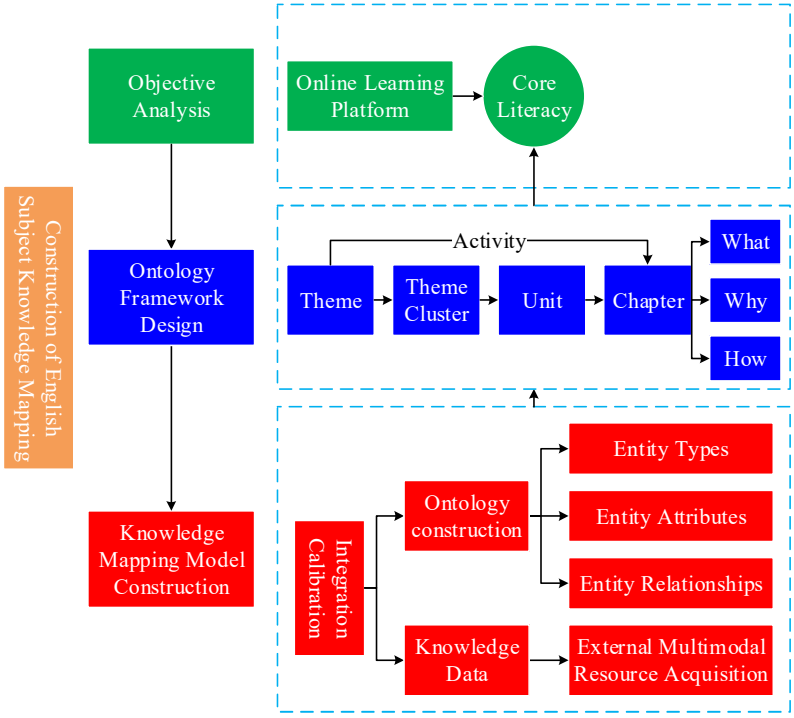

Figure 1: Construction process of English subject knowledge map

II. B.Data preparation

After clarifying the goal and structure of the construction of English subject knowledge mapping, its dynamic characterization needs to be realized through specific data. This section will elaborate the data preparation process, including exercise annotation, adjacency matrix generation and feature vector extraction, to provide a data basis for the subsequent knowledge transfer method.

When generating a personalized knowledge map for each student, firstly, a set of the same exercises is generated for each student based on the original knowledge map, which covers all the knowledge points and the number of exercises is 100, through which we can obtain the mastery of different students on the knowledge points. After that, according to the student's doing this part of the knowledge map of the exercises labeled, students do the right questions as mastered, the label is set to 1, students do the wrong questions as not mastered, the label is set to 0. After that, for the entire knowledge map to generate the corresponding adjacency matrix, and the use of word2vec to generate the dimensionality of 512 feature vectors for each exercise, and then the entire exercise dataset will be cut into training set and test set, where the number of training set is 1037 and the number of test set is 40, and all the exercises in the test set have labels, and some of the exercises in the training set have labels.

The computer needs to store the graph data as an adjacency matrix or an adjacency table. An adjacency table is a combination of sequential and chained storage, where each node in the graph is first numbered and the number corresponding to the neighboring nodes of each node is stored in a chained table, after which all the nodes are stored in sequential order by number in a sequential table. The weights on the edges between the nodes are stored in the chained table when using the neighbor table to store a weighted graph, and no data is stored when storing an unweighted graph. When storing an undirected graph using an adjacency table, each node connected to the node is stored in the chain table corresponding to the node, and when storing a directed graph, only the nodes pointing from the node are stored in the chain table corresponding to the node. The corresponding neighbor table for the undirected graph is shown in Figure 2.

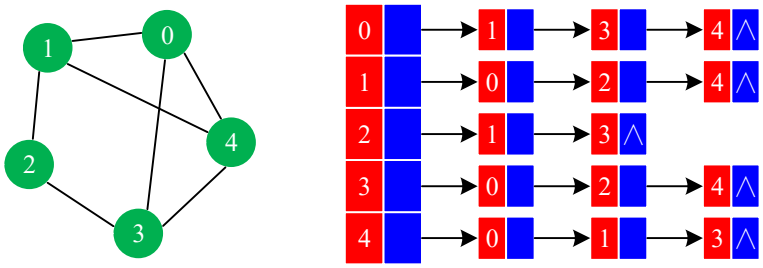

Figure 2: Undirected graphs correspond to adjacency lists

When using the adjacency matrix to store the graph data it is also necessary to number each node in the graph, after which the subscripts of the matrix are used to represent the nodes in the graph, and the values in the matrix are used to represent the relationships between the nodes. For an unweighted graph, when two nodes are connected by an edge, the value of the corresponding position in the matrix is set to 1, and vice versa to 0. For an undirected weighted graph, when two nodes are connected by an edge, the value of the corresponding position in the matrix is set to the weight on the edge, and vice versa to 0. For an undirected graph, if the nodes are connected by an edge, the corresponding adjacency matrices in X are X_{ij} and X_{ji}. and X_{ji} have the same value, i.e., the corresponding adjacency matrix of the undirected graph is a symmetric array In this paper, the knowledge graph generated is an undirected graph composed of exercise nodes, and its generated adjacency matrix is a symmetric array.

When generating the knowledge graph, the corresponding adjacency matrix can be generated directly by using the numpy extension package of python and according to the definition of the adjacency matrix. numpy is a scientific computing tool that supports the python programming language and can be easily used for matrix arithmetic and storage. In generating the adjacency matrix with numpy, first define a two-dimensional matrix of size 9673*9673 with all values 0, where 9673 is the number of nodes in the knowledge graph. When the similarity between the two exercise nodes is higher than the threshold defined in this paper, these two nodes are connected with an edge in the knowledge graph, i.e., the value of the corresponding position in the adjacency matrix is to be set to 1. While generating edges for the two nodes in the knowledge graph, the value of the corresponding position in the adjacency matrix is changed, and therefore its adjacency matrix is also generated when generating the knowledge graph.

When generating the feature vectors corresponding to the exercise data in the test set, the feature vectors corresponding to the exercises in the training set, and the feature vectors corresponding to the labeled exercise data in the training set, we need to use word2vec to generate the feature vectors for each exercise, and the dimension of the feature vectors in this paper is set to 512. In this paper, we choose the jieba lexical tool to perform the operation of the parsing of the lexical words. After obtaining all the words involved in the exercise data, the stem of the exercise and the English words involved are passed to the word2vec model for training, then word vectors can be generated for each English word, and after obtaining the word vectors corresponding to each word, the word vectors of all the words contained in the stem of the exercise are summed up and divided by the number of words contained in the stem, n, then the sentence vectors corresponding to the stem can be obtained. Sentence vector SQ, as in equation (1). Where SQ_k denotes the value of the k th dimension of the sentence vector SQ, and W_{ik} denotes the value of the k th dimension of the i th word vector in the stem.

$$\begin{aligned} SQ &= \left(SQ_1, SQ_k \cdots SQ_{512}\right) \\ &= \left(\frac{\sum_{i=1}^{n} W_{i1}}{n}, \frac{\sum_{i=1}^{n} W_{ik}}{n} \cdots \frac{\sum_{i=1}^{n} W_{i512}}{n}\right) \end{aligned} \tag{1}$$

Similarly, the parsing and the Chinese words involved are passed to the word2vec model for training, then a word vector can be generated for each Chinese word, and then the corresponding sentence vector of the parsing can be obtained in the same way, and then the sentence vector corresponding to the stem and the sentence vector corresponding to the parsing can be summed up and divided by 2 to obtain the feature vector ST of the exercise, as shown in Equation (2). Where ST_k denotes the value of the k th dimension of the feature vector, SQ_k denotes the value of the k th dimension of the sentence vector, and SA_k denotes the value of the k th dimension of the vector corresponding to the parsing of the exercise.

$$ST = (ST_1, ST_k \cdots ST_{512}) = \left(\frac{SQ_1 + SA_1}{2}, \frac{SQ_k + SA_k}{2} \cdots \frac{SQ_{512} + SA_{512}}{2} \right) \quad (2)$$

After obtaining the feature vectors of all exercises, the above three files are obtained separately according to the test set and the training set divided before, and the adjacency matrix corresponding to the knowledge graph is the same, the data type of these three files in the computer memory is ndarray, and the data type needs to be converted into scipy.sparse.csr.csr_matrix by using the scipy tool, and then serialized and saved to the local by using the pickle tool. The suffix of the feature vector file corresponding to the exercise data in the test set is named .tx, the suffix of the feature vector file corresponding to the exercise set is named .allx, and the suffix of the feature vector file corresponding to the exercise data with labels in the training set is named.

II. C.Joint Knowledge Migration Method Based on Local Linear Embedding

Based on the generated knowledge map and feature vectors, how to realize cross-domain knowledge migration becomes critical. In this section, we propose a local linear embedding-driven joint migration method, HLLEJKT, which effectively aligns the multilingual representation space by optimizing the mapping matrix with kernelized extensions, and provides technical guarantee for personalized learning support.

II. C. 1) Joint Knowledge Migration Method Based on Similar Localized Linear Embedding

For the same kind of local linear embedded joint knowledge transfer (HLLEJKT) method, because the target domain contains a small number of labeled samples, and LLEJKT does not take into account the labeling information of the target domain in the learning process, in order to better learn the data distribution of the source and target domains, in this paper, we hope to make use of the labeling information of the samples of the target domain, so that the sample intraclass distance is small, and the distance between classes is large. At this time, only need to consider the information of the same kind of data each node k near neighbors, and in the low-dimensional space to maintain the original high-dimensional space of the same kind of samples within the neighborhood of the linear relationship, HLLEJKT algorithm steps are shown in Figure 3.

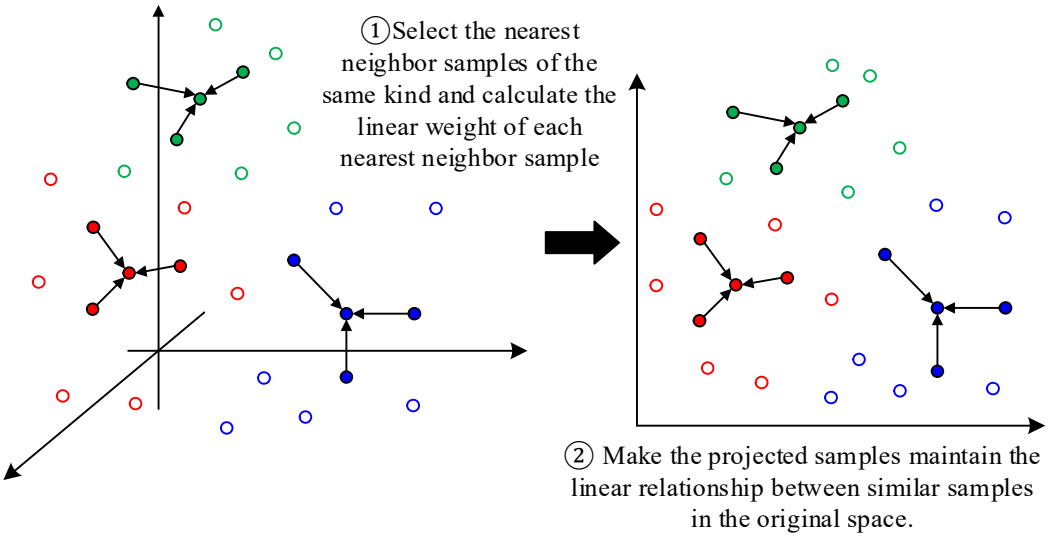

Figure 3: HLLEJKT algorithm steps

Therefore, for the target domain data, this paper wishes to find a mapping matrix B such that the mapped target domain data still maintains its linear relationship within the neighborhood between similar samples in the original high-dimensional space. Then the optimization objective of maintaining the local linear relationship between the same kind becomes:

$$\begin{aligned}&\min_w \sum_{c=1}^{C}\sum_{i=1}^{N_t^{(c)}}\left\|z_i^{(c)}-\sum_{j=1}^{k}w_{ji}^{(c)}z_{ji}^{(c)}\right\|^2\\&s.t.\ \sum_{j=1}^{k}w_{ji}^{(c)}=1\end{aligned}\tag{3}$$

$$\begin{aligned}&\min_{Z'} \sum_{c=1}^{C}\sum_{i=1}^{N_t^{(c)}}\left\|z_i'^{(c)}-\sum_{j=1}^{k}w_{ji}^{(c)}z_{ji}'^{(c)}\right\|^2\\&s.t.\ \sum_{i=1}^{N_t^{(c)}}z_i'^{(c)}=0\\&\sum_{i=1}^{N_t^{(c)}}z_i'^{(c)}\left(z_i'^{(c)}\right)^T-N_t^{(c)}I=0\end{aligned}\tag{4}$$

where $N_t^{(c)}$ denotes the number of samples belonging to the c th class of data in the target domain, $z_i^{(c)}$ denotes the i th sample of the c th class of data, $z_{ji}^{(c)}$ denotes the j th nearest neighbor of the sample $z_i^{(c)}$, and $w_{ji}^{(c)}$ denotes the weight of the $z_i^{(c)}$ of the j th nearest neighbor, $z_i'^{(c)}$ denotes the i th sample belonging to the c th class of data in the projected target domain, and $z_{ji}'^{(c)}$ denotes the j th nearest neighbor of the sample $z_i^{(c)}$ in the original space after projection. Then the final optimization objective becomes:

$$\begin{aligned}&\min_{A,B}\alpha\ tr\left(A^TS_wA\right)+\beta\sum_{c=1}^{C}tr\left(B^TZ^{(c)}V^{(c)}\left(Z^{(c)}\right)^TB\right)\\&+D'_{S,T}+\rho\left(\|B-A\|_F^2+\|B\|_F^2\right)\\&s.t.\ \sum_{c=1}^{C}tr\left(B^TZ^{(c)}\left(Z^{(c)}\right)^TB-N_t^{(c)}I\right)=0\\&tr\left(A^TS_bA-I\right)=0\end{aligned}\tag{5}$$

where $Z^{(c)}$ denotes the data belonging to the c th class in the target domain, and $V^{(c)}=\left(I-W^{(c)}\right)\left(I-W^{(c)}\right)^T$, where $W^{(c)}$ is a sparse matrix.

II. C. 2) Nuclearization analysis

To further enhance the migration effect under nonlinear data, this section introduces kernelization analysis to extend HLLEJKT to regenerative Hilbert space, which enhances the model's ability of modeling complex semantic relations through kernel function mapping, and ultimately achieves a performance breakthrough in cross-language tasks.

In the regenerated Hilbert space, this paper can obtain the nonlinear LLEJKT and HLLEJKT by kernelization, i.e., KLLEJKT and KHLLEJKT. Here this paper focuses on the KLLEJKT method. In this paper, we make the kernel function $\phi: x\mapsto\phi(x)$, defined as $\Phi(S)=\left[\phi(x_1),\cdots,\phi(x_{N_s}),\phi(z_1),\cdots,\phi(z_{N_t})\right]\in\mathbb{R}^{D\times N}$, where $N=N_s+N_t$. In this paper, we use $A=\Phi(S)A$, $B=\Phi(S)B$ to denuclearize LLEJKT, where $S=[X,Z]$, $A\in\mathbb{R}^{N\times d}$ and $B\in\mathbb{R}^{N\times d}$ are two projection matrices to be optimized.

In this paper, we first make $K_S=\Phi(S)^T\Phi(X)$ and $K_T=\Phi(T)^T\Phi(Z)$. Then all x in the source domain and all z in the target domain are replaced by $\phi(x)$ and $\phi(z)$, i.e., $x\to\phi(x)$, $z\to\phi(z)$, and the objective function can be expressed as in the above derivation:

$$\begin{aligned}&\min_{A,B}\alpha\ tr\left(A^TS_wA\right)+\beta\ tr\left(B^TK_TVK_T^TB\right)\\&+\left\|M_S^TK_S^TA-M_T^TK_T^TB\right\|_F^2+\rho\left(W^TUW\right)\\&s.t.\ tr\left(B^TK_TK_T^TB-N_tI\right)=0\\&tr\left(A^TS_bA-I\right)=0\end{aligned}\tag{6}$$

where $S_w = \sum_{c=1}^{C} K_S^{(c)} H_s^{(c)} \left(K_S^{(c)}\right)^T$, and $K_S^{(c)}$ denotes the portion of K_S that belongs to the c th class only, $H_s^{(c)} = I_{N_s^{(c)}} - \frac{1}{N_s^{(c)}} 1_{N_s^{(c)}} \left(1_{N_s^{(c)}}\right)^T$ is the center matrix, and for S_b, μ_c denotes the mean of $K_S^{(c)}$ and μ denotes the mean of $K = [K_S, K_T]$.

$V = (I-W)(I-W)^T$, W is a sparse matrix consisting of w_i. $w_i = \frac{s_i^{-1} 1_k}{1_k^T s_i^{-1} 1_k}$, where the kernelized $s_i = K_{T(i)} 1_k 1_k^T - 2 1_k 1_k^T K_{T(i)} + K_{TZ}$.

where

$$K_{T(i)} = \phi(z_i^T)\phi(z_i) \tag{7}$$

$$\begin{aligned} K_{T(i)} &= \left[K_{T(i,1)}\ K_{T(i,2)} \cdots K_{T(i,j)} \cdots K_{T(i,k)}\right] \\ K_{T(i,j)} &= \phi\left(z_i^T\right)\phi(z_{ji}) \end{aligned} \tag{8}$$

$$\begin{aligned} K_{TZ} &= \begin{bmatrix} K_{(T,i)(1,1)} & K_{(T,i)(1,2)} & \cdots & K_{(T,i)(1,k)} \\ K_{(T,i)(2,1)} & K_{(T,j)(2,2)} & \cdots & K_{(T,i)(2,k)} \\ \vdots & \vdots & \ddots & \vdots \\ K_{(T,i)(k,1)} & K_{(T,i)(k,2)} & \cdots & K_{(T,i)(k,k)} \end{bmatrix} \\ K_{(T,i)(j,k)} &= \phi\left(z_{ji}^T\right)\phi(z_{ki}) \end{aligned} \tag{9}$$

Similarly, the objective function of KHLLEJKT is:

$$\begin{aligned} &\min_{A,B} \alpha\, tr\left(A^T S_w A\right) + \beta \sum_{c=1}^{C} tr\left(B^T K_T^{(c)} V^{(c)} \left(K_T^{(c)}\right)^T B\right) \\ &+ \left\| M_S^T K_S^T A - M_T^T K_T^T B \right\|_F^2 + \rho\left(W^T U W\right) \\ &s.t. \sum_{c=1}^{C} tr\left(B^T K_T^{(c)} (K_T^{(c)})^T B - N_t^{(c)} I\right) = 0 \\ &tr\left(A^T S_b A - I\right) = 0 \end{aligned} \tag{10}$$

where $S_w = \sum_{c=1}^{C} K_S^{(c)} H_s^{(c)} (K_S^{(c)})^T$, $K_S^{(c)}$ represents the part of K_S that belongs only to class c, and $K_T^{(c)}$ represents the part of K_T that belongs only to class i, $H_s^{(c)} = I_{N_s^{(c)}} - \frac{1}{N_s^{(c)}} 1_{N_s^{(c)}} \left(1_{N_s^{(c)}}\right)^T$ is the central matrix, for S_b, μ_c is the mean of $K_s^{(c)}$, μ is $K = [K_S, K_T]$. Replace I with K to get U.

$V^{(c)} = (I - W^{(c)})(I - W^{(c)})^T$, where $W^{(c)}$ is a sparse matrix consisting of $w_i^{(c)}$. $w_i = \frac{(S_i^{(c)})^{-1} 1_k}{1_k^T (S_i^{(c)})^{-1} 1_k}$, where the kernelized $S_i^{(c)} = K_{T(i)}^{(c)} 1_k 1_k^T - 2 1_k 1_k^T K_{T(i)}^{(c)} + K_{TZ}^{(c)}$.

where

$$K_{T(i)}^{(c)} = \phi\left((z_i^{(c)})^T\right)\phi\left(z_i^{(c)}\right) \tag{11}$$

$$\begin{aligned} K_{T(i)}^{(c)} &= \left[K_{T(i,1)}^{(c)}\ K_{T(i,2)}^{(c)} \cdots K_{T(i,j)}^{(c)} \cdots K_{T(i,k)}^{(c)}\right] \\ K_{T(i,j)}^{(c)} &= \phi\left((z_i^{(c)})^T\right)\phi\left(z_{ji}^{(c)}\right) \end{aligned} \tag{12}$$

$$K_{TZ}^{(c)} = \begin{bmatrix} K_{(T,i)(1,1)}^{(c)} & K_{(T,i)(1,2)}^{(c)} & \cdots & K_{(T,i)(1,k)}^{(c)} \\ K_{(T,i)(2,1)}^{(c)} & K_{(T,j)(2,2)}^{(c)} & \cdots & K_{(T,i)(2,k)}^{(c)} \\ \vdots & \vdots & \ddots & \vdots \\ K_{(T,i)(k,1)}^{(c)} & K_{(T,i)(k,2)}^{(c)} & \cdots & K_{(T,i)(k,k)}^{(c)} \end{bmatrix} \quad (13)$$

$$K_{(T,i)(j,k)}^{(c)} = \phi\left((z_{ji}^{(c)})^T\right)\phi\left(z_{ki}^{(c)}\right)$$

III. Experiments and Performance Analysis of Multilingual Knowledge Migration Based on HLLEJKT

After completing the design of the construction and migration method for English subject knowledge mapping, its effectiveness needs to be verified through experiments. Chapter 3 will systematically evaluate the performance of the HLLEJKT method based on multilingual datasets and benchmark tasks, and deeply analyze its advantages and limitations in different language scenarios.

III. A. Test data

III. A. 1) Data sets

In this study, the data from the "2010 People's Daily Corpus Entity Recognition Annotation Set" is migrated to English corpus data with NER tags using a placeholder approach. In order to test the performance of knowledge migration for bi-lingual and multi-lingual cross-language models, both Spanish and Arabic are selected for data migration, because these two languages are important corpus components for CINO secondary pre-training, and can be better for knowledge migration. Spanish has a natural separator, and data migration is carried out as in English; since Arabic has no separator, it needs to be segmented after data migration, and the TIP-LAS tool is used to segment Arabic, which is processed to synthesize Arabic NER labeled data. In order to ensure the quality and reliability of the machine translation corpus, a number of linguistics graduate students in English, Spanish and Arabic were recruited to evaluate the quality of the data, and the evaluation results meet the requirements of the NER task, and the specific data information is shown in Table 1, in which Train+MR denotes the experimental data after entity enhancement is carried out, Train denotes the training set, Dev denotes the calibration set, and Test denotes the test set.

Table 1: Information on the experimental corpus

Test data	Data type	Sentence number	Number of names	Number of place names	Number of organization names
Chinese	Train+MR	30182	14292	27363	16394
	Train	26384	8642	18264	10385
	Dev	3721	1022	2176	925
	Test	5018	2016	4092	2438
English	Train+MR	29784	13874	26935	15938
	Train	23254	8203	18027	10342
	Dev	3417	984	2078	911
	Test	4723	1940	3918	2381
Spanish	Train+MR	29781	13899	26915	15974
	Train	23293	8246	18074	10355
	Dev	3389	1016	2116	898
	Test	4694	1974	3899	2385
Arabic	Train+MR	29768	13934	27003	15870
	Train	23330	8173	18061	10288
	Dev	3376	1005	2155	947
	Test	4711	1897	3931	2456

III. A. 2) Description of tasks

Using the HLLEJKT method to migrate the knowledge from the English contrastive learning model to the multilingual pre-trained language model, a total of four experiments on knowledge migration in the English-English, Chinese, Spanish and Arabic directions were conducted. The migration in the English-English direction does not require a parallel corpus, so this experiment collects sentences from the above dataset as training data. For the other 3

directions, the parallel corpus of English-to-that-direction is used as training data, and 20,000 data are collected for each direction.

This experiment evaluates the performance of the model on the NLI task and the STS task. In the NLI task, the input to the model is a pair of sentences, called premise and hypothesis, respectively, and the model is asked to judge the category of logical relationship (implication, neutrality, contradiction) between the 2 sentences, with accuracy as the evaluation index. Specifically, the NLI benchmark chosen for this experiment is XNLI, which is an NLI benchmark involving 15 languages, each with its own validation and test sets, and which provides only English training data; however, it provides machine-translated versions of the English training data in other languages in the related resources, which can be used for model fine-tuning. The experiments were evaluated only on the English, Arabic, Spanish and Chinese test sets of XNLI.

In the STS task, the input to the model is a pair of sentences and the output is the cosine similarity of the 2 sentences, with the Spearman correlation score between the model output and the labels as the evaluation metric. Specifically, the STS 2020 benchmark was chosen for the experiments.STS 2020 is a multilingual STS benchmark for English, Spanish, and Arabic that contains both monolingual and cross-language tasks, and the experiments were evaluated on its monolingual tasks (English, Spanish, and Arabic) only.

III. A. 3) Experimental setup

The experiment aligns the subrepresentation spaces occupied by each of English, Spanish, Chinese and Arabic in the representation space of the multilingual pre-trained language model with that of the English contrastive learning model by using the HLLEJKT method, so as to migrate the knowledge of the English contrastive learning model to the multilingual pre-trained language model simultaneously along the four directions, namely English-English, Chinese, Spanish and Arabic. Since local structural coding needs to be computed for each language during the training process, the current small batch must contain training data in only one direction, and this paper uses an alternating training method to achieve this purpose.

During training, the batch_size is set to 64, the number of training rounds is 5, the optimizer is AdamW, the initial learning rate is 5×10-6, the learning rate scheduling strategy is preheating linear decay, the number of preheating steps is 1000, and the model parameter that scores the highest on the validation set is retained by validating it at intervals of 125 steps.

After the knowledge migration is completed, the multilingual pretrained language model in HLLEJKT is removed and the quality of the sentence representations produced by it is evaluated, thus verifying that the knowledge from the English contrastive learning model is successfully migrated to the multilingual pretrained language model. The evaluation freezes the parameters of the multilingual pre-trained language model and uses it only as an encoder for acquiring sentence representations.

When evaluating on the XNLI benchmark, firstly the model parameters are frozen and secondly a classifier containing one hidden layer (a fully connected layer of 512 neurons) is added after the model. This classifier is trained on the corresponding language training data provided by XNLI with a learning rate of 0.01, an optimizer of Stochastic Gradient Descent (SGD), a batch_size of 128, and a total of 2 rounds of training.

When evaluated on the STS 2020 benchmark, the model parameters were still frozen, and the input 2 sentences were first encoded, then the cosine similarity of the 2 sentence representations was directly used as the model predictions to compute the Spearman's correlation coefficients between them and the labels, without using any training data for targeted fine-tuning.

III. A. 4) Contrasting models

The following models are selected for comparison in this experiment:

(1) mBERT: Same architecture as BERT, pre-training with masked language modeling (MLM) on a large-scale corpus containing 104 languages, with model parameters shared across all languages.

(2) Align: aligning sentence representations in a teacher-student architecture, thus enabling cross-lingual knowledge transfer. The student model is a bilingual encoder for languages A and B, and the teacher model is a monolingual encoder for language A. The models are trained on parallel corpora of A and B using the mean square error as a loss function, so that the encoding of sentences in the 2 languages A and B by the student model is simultaneously aligned with the encoding of sentences in language A by the teacher model, thus transferring knowledge from the monolingual teacher model to the multilingual student model. In this paper, we denote this method as Align, which indicates that the method is based on representational alignment for cross-language knowledge transfer.

(3) mSimCSE: It fine-tunes the cross-language pre-trained language model XLM-R with a contrastive learning approach on English NLI data, and all languages share the parameters of XLM-R, and the knowledge learned by

the model from the English NLI data can be generalized to other languages through the parameters of XLM-R. The trained model achieves a very excellent performance.

(4) LLEJKT: In this setting, the data enhancement strategy is not used, and a comparison is made with a similar HLLEJKT to see whether the data enhancement using Dropout brings any performance gain.

III. B. Data Enhancement Performance Testing

Based on the clarification of the experimental dataset and task settings, this section will focus on the specific performance of the HLLEJKT method in cross-language knowledge transfer, revealing its breakthrough enhancement over traditional methods through comparative experiments and data enhancement tests.

III. B. 1) Comparative Performance Analysis of Cross-Language Benchmarking Tasks

Table 2 shows the experimental results on the XNLI and STS 2020 benchmarks.

Table 2: Accuracy comparison between XNLI and STS 2020 benchmarks/%

	Model	English	Chinese	Spanish	Arabia	Average
XNLI	mBERT	52.07	46.67	50.98	47.03	49.19
	Align	67.48	62.97	65.78	64.57	65.20
	mSimCSE	70.61	65.31	69.07	66.44	67.86
	LLEJKT	73.31	68.59	72.12	70.09	71.03
	HLLEJKT	77.66	72.18	77.45	73.22	75.13
STS 2020	mBERT	40.94	32.64	38.29	37.29	37.29
	Align	82.91	67.65	78.81	69.99	74.84
	mSimCSE	86.47	70.25	83.43	76.64	79.20
	LLEJKT	87.88	72.11	86.24	69.34	78.89
	HLLEJKT	90.63	80.24	88.59	83.53	85.75

HLLEJKT achieved the best performance on both 2 benchmarks. Compared to mBERT, the models trained using the HLLEJKT method achieved a significant performance improvement of 55.69 and 124 percentage points on average on the XNLI and STS 2020 benchmark datasets, respectively, which demonstrates the effectiveness of the local linear embedding-based joint knowledge migration method for English constructed in this paper.

Compared to Align, HLLEJKT achieves a performance improvement of 13.40 and 19.35 percentage points on the XNLI and STS 2020 benchmarks, respectively. This indicates that when performing cross-language knowledge migration for contrastive learning models, the approach based on aligning the structure of the representation space of different languages is superior to the approach of directly aligning the representations, which side-steps the fact that the knowledge of the contrastive learning models exists in the structure of the representation space rather than in the representations themselves.

III. B. 2) Multilingual Entity Recognition Performance Evaluation

In this study, the accuracy A, precision P, recall R and F1 value are used as the evaluation indexes of the model, and the formulas are calculated as follows, respectively:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{14}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{15}$$

$$F1 = \frac{2PR}{P + R} \tag{16}$$

where TP is the number of correctly recognized entities, FP is the number of incorrectly recognized entities, and FN is the number of unrecognized entities.

The traditional method of labeling entities and relations is used, and the BIO scheme is used for entity labeling, and entities with relations are classified and labeled, and comparative experiments are carried out on the above models, and the test results of the models are shown in Fig. 4.

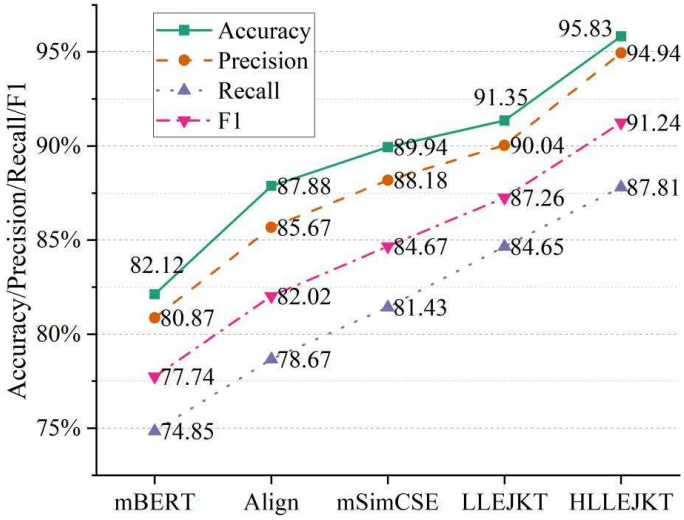

Figure 4: Performance comparison of entity and relationship extraction models

Figure 4 illustrates the performance metrics of different models in the entity recognition task, including accuracy (A), precision (P), recall (R) and F1 value.The HLLEJKT model performs optimally, with an accuracy of 95.83%, a precision of 94.94% and an F1 value of 91.24%, which are significantly higher than the other models. Specifically, HLLEJKT compares with mBERT: the accuracy rate is improved by 16.70% and the F1 value is improved by 17.37%, which indicates that the knowledge migration method based on local linear embedding can effectively enhance the model's ability of recognizing the entity boundaries.HLLEJKT compares with Align: the F1 value is improved by 11.23%, which reflects that the migration strategy combining with the labeling information is better than the pure representation alignment method.

III. C. Error analysis

Although HLLEJKT performs well in the benchmark task, it still needs to be explored for potential flaws through error analysis. In this section, we will parse the performance differences of the model in different language directions based on the confusion matrix and misclassification cases to provide a basis for subsequent optimization.

Error analysis is one of the important methods for training and fine-tuning models. When the model performance is much worse than expected, error analysis can provide insight into the strengths and weaknesses of the model, and is a powerful tool to understand the strengths and weaknesses of the model.

The input tokens are grouped, and the number, mean, and sum of each token are aggregated to remit a confusion matrix of cross-language entity tokens in four directions as shown in Figure 5.

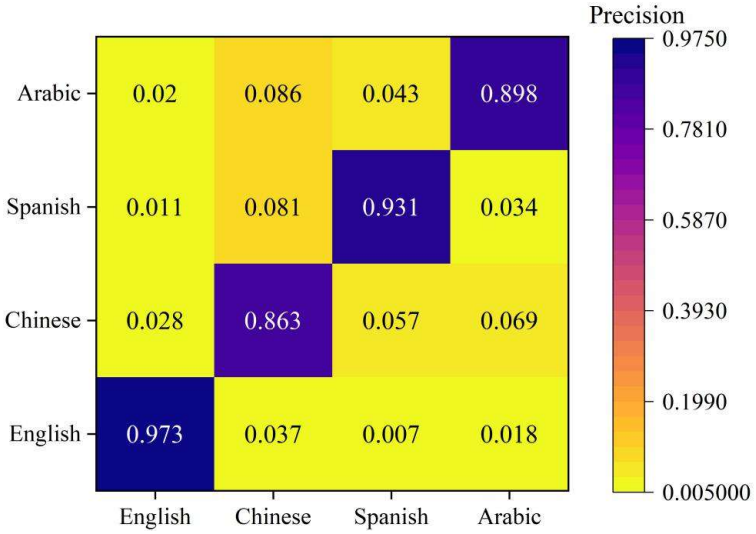

Figure 5: Confusion matrix

Figure 5 demonstrates the recognition accuracy of entity labeling under different language orientations through the confusion matrix: English → English: the diagonal accuracy is the highest at 97.3%, indicating that the model has the strongest ability to recognize entities in the same language. Chinese → other languages: Chinese entities are misclassified as Spanish and Arabic at a higher rate of 8.1% and 8.6% respectively, which may be related to the smaller morphological differences between languages. Spanish and Arabic: the recognition accuracy within Spanish, at 93.1%, is higher than that of Arabic, at 89.8%, probably due to the lower generalization ability of the model caused by the higher complexity of Arabic participles. The percentage of English entities misclassified as other languages was the lowest, both ≤ 2.8%, verifying the dominant role of English as the source language in knowledge transfer.

IV. Knowledge Graph-based Application of English Teaching Resources Retrieval in Colleges and Universities

After verifying the theoretical superiority of the HLLEJKT method, it becomes a natural extension to see how the knowledge mapping technology can be transformed into a practical teaching tool. Chapter 4 develops an intelligent teaching resource retrieval system based on the English knowledge graph of the local linear embedding algorithm constructed in the previous section, and verifies the application value of the theoretical results through empirical research.

IV. A. Comparative Experiment on Retrieval Accuracy of Multilingual Teaching Resources

The Knowledge Graph Intelligent Retrieval Method of College English Teaching Resources Based on Locality Embedding Algorithm proposed in the article is set as the experimental group, and the Fuzzy Retrieval Method of Teaching Resources, the Resource Retrieval Method Based on Similarity Matching, and the Retrieval Method Based on Ontology are set as the Control Group 1, Control Group 2, and Control Group 3, respectively, and the intelligent retrieval results of the four methods are compared.

MATLAB software was used to simulate the whole process of intelligent retrieval of the three methods, and the accuracy of intelligent retrieval of teaching resources of the four entity categories in the dataset was determined respectively. To ensure the objectivity of the experimental results, 10 experiments were conducted, and the results of the comparison of the intelligent retrieval accuracy of teaching resources are shown in Figure 6.

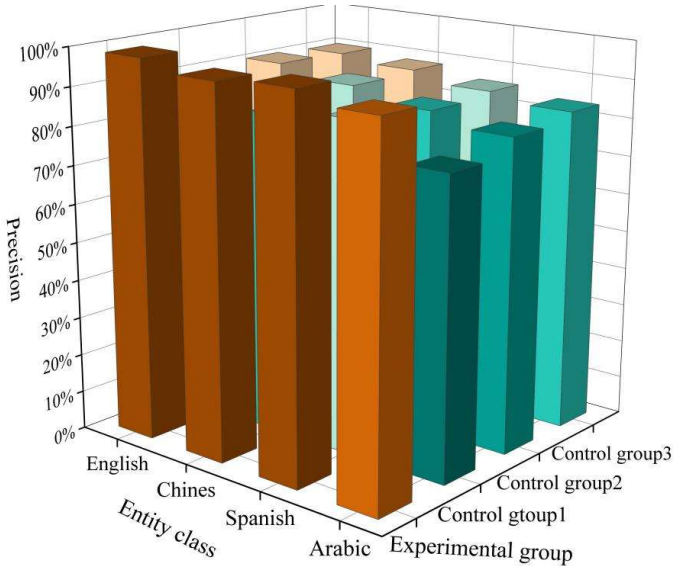

Figure 6: Comparison of the precision of intelligent retrieval of teaching resources

The HLLEJKT-based retrieval method achieves significant advantages in all four language resource retrieval. The accuracy rate of English resource retrieval reaches 98.47%, which is 11.61 percentage points higher than the fuzzy retrieval of control group 1. In the low-resource language Arabic scenario, the accuracy rate of the experimental group is 94.56%, which is 10.89 percentage points higher than the advantage of ontology retrieval of control group 3. The average multilingual accuracy rate of 96.55% is 11.5, 10.78, and 9.94 percentage points higher than the 85.05%, 85.77%, and 86.61% of the three control groups, respectively. It is worth noting that the Spanish retrieval accuracy shows a special distribution, with 97.33% in the experimental group significantly better than 84.48% in the

control group 2 based on similarity matching, which verifies the knowledge graph's strong ability to capture linguistic morphological features.

IV. B. Validation of Knowledge Graph Retrieval Path Optimization Effects

Accuracy validation only reflects the basic performance of the system, and retrieval efficiency is equally critical in real-world applications. This section further reveals the efficiency improvement brought by the optimization of knowledge graph structure through path length analysis.

Retrieval is performed under this knowledge graph, and the retrieval path lengths under different methods are compared. The results of the average retrieval path lengths under different retrieval methods are shown in Figure 7.

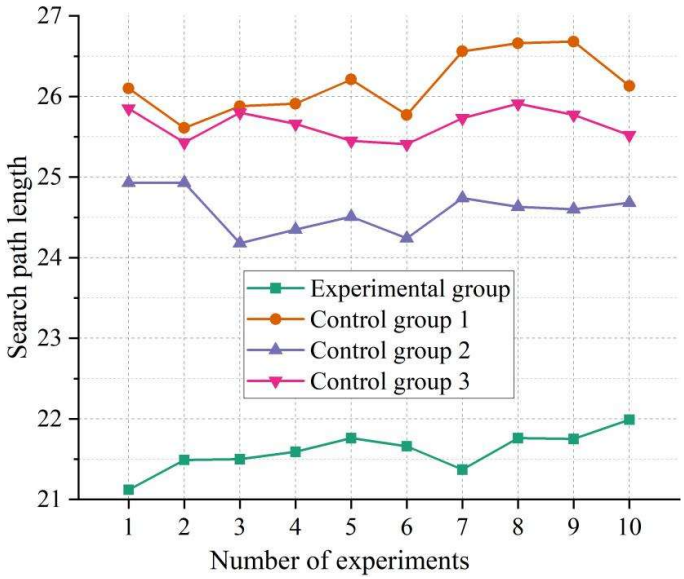

Figure 7: The average search path length under different search methods

From the above results, it can be seen that in the process of retrieving English education network resources in colleges and universities under 20 nodes, the retrieval path using the knowledge graph-based remote multimedia education network resources retrieval method designed in this paper is shorter, with an average of 21.60, and the traditional fuzzy retrieval path of teaching resources has an average of 26.15; the similarity-matching-based resources retrieval path has an average of 24.58; and the retrieval path based on the Ontology-based retrieval method retrieval path is 25.65 on average, which verifies the effectiveness of this paper's method.

V. Conclusion

In this study, we construct a dynamic English knowledge graph by local linear embedding algorithm and propose HLLEJKT joint knowledge transfer method, which achieves significant results in multilingual education scenarios.

HLLEJKT achieves 77.66%, 72.18%, 77.45%, and 73.22% accuracy in the English, Chinese, Spanish, and Arabic tasks of the XNLI benchmark, respectively, which is an average improvement of 25.94 percentage points over the baseline model mBERT; and the accuracy in the STS 2020 task is 90.63% (English) and 83.53% (Arabic), validating the advantages of the kernelization extension for modeling nonlinear semantic relations.

In the Entity Tagging task, the F1 value of HLLEJKT reaches 91.24%, which is an improvement of 11.23% over the Align method, especially showing strong robustness in morphologically complex languages such as English recognized as 93.1% for Spanish and 89.8% for Arabic.

The graph-based intelligent retrieval system achieves 98.47% retrieval accuracy for English resources, and the English retrieval path length is optimized to 21.60, while the traditional fuzzy retrieval path for teaching resources averages 26.15, which verifies the enhancement of the local linear structure on the association of multilingual resources.

References

[1] Wei, Z. (2024). Navigating digital learning landscapes: unveiling the interplay between learning behaviors, digital literacy, and educational outcomes. Journal of the Knowledge Economy, 15(3), 10516-10546.

[2] Yang, X. (2023, August). Construction and Application of Digital College English Teaching Resources on the Basis of Data Mining. In EAI International Conference, BigIoT-EDU (pp. 146-153). Cham: Springer Nature Switzerland.
[3] Xu, C. (2025). Intelligent recommendation method for digital teaching resources of online courses based on knowledge graph. International Journal of Continuing Engineering Education and Life Long Learning, 35(1-2), 62-76.
[4] Song, Y., Sun, P., Liu, H., Li, Z., Song, W., Xiao, Y., & Zhou, X. (2024). Scene-driven multimodal knowledge graph construction for embodied ai. IEEE Transactions on Knowledge and Data Engineering.
[5] Liao, X., Chen, C., Wang, Z., Liu, Y., Wang, T., & Cheng, L. (2025). Large language model assisted fine-grained knowledge graph construction for robotic fault diagnosis. Advanced Engineering Informatics, 65, 103134.
[6] Yang, Y., Peng, X., Chen, M., & Liu, S. (2025). An explainable graph-based course recommendation model based on multiple interest factors. Expert Systems with Applications, 264, 125889.
[7] Liu, S., Xu, L., Liu, Y., & Kolmanič, S. (2025). Dual-view embedding for hyper-relational knowledge graphs with hierarchical structure. Journal of Intelligent Information Systems, 1-18.
[8] Nguyen Thi Kim, L., Nguyen Hoang, S., & Nguyen, H. N. (2025). Interweaving academic insights: advancing university knowledge management through a strategic data fabric framework. Digital Library Perspectives, 41(1), 21-44.
[9] Xia, X., & Qi, W. (2025). Learning behaviour prediction and multi-task recommendation based on a knowledge graph in MOOCs. Technology, Pedagogy and Education, 1-24.
[10] Gustafson, J. R. D., Jhajj, G., Zhang, X., & Lin, F. O. (2025). Enhancing Project-Based Learning With a GenAI Tool Based on Retrieval: Augmented Generation and Knowledge Graphs. In AI Applications and Strategies in Teacher Education (pp. 161-194). IGI Global.
[11] Huang, Y., & Zhu, J. (2021). A personalized English learning material recommendation system based on knowledge graph. International Journal of Emerging Technologies in Learning (Online), 16(11), 160.
[12] Wang, L. (2022). An improved knowledge graph question answering system for english teaching. Mobile Information Systems, 2022(1), 3401074.
[13] Wu, Z., & Jia, F. (2022). Construction and application of a major-specific knowledge graph based on big data in education. International Journal of Emerging Technologies in Learning (iJET), 17(7), 64-79.
[14] Hu, L., Chen, Y., & Chen, L. (2025). A study on the impact of diverse evaluation system on college students' sense of achievement in English learning: An empirical research based on the knowledge graphs of College English. Education and Information Technologies, 1-30.