International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-10

https://doi.org/10.70917/ijcisim-2026-1114
Article

Analyzing Regional Variation of English Dialects in Central Asia Using Text Similarity Algorithm - An Exploration of Similarity Measure and Clustering Features

Yu Zhao [1] and Yulong Liang [1,*]

[1] YiLi Normal University, Yining, Xinjiang, 835000, China
* Correspondence author: yn_lyl@163.com

Abstract: With the acceleration of globalization, English has become a universal language, but its localized use in different regions has given rise to numerous regional variations with systematicity. The lack of large-scale annotated corpus for English in Central Asia makes traditional supervised categorization difficult to implement, for this reason, this paper reveals regional text clustering based on context and mapping relationships through text similarity algorithm. Then, the traditional Topic Frequency-Inverse Document Frequency (TFIDF) is improved from the perspective of inter- and intra-category discretization, and the improved ITFIDF method is used to calculate the feature weights. The results show that in the self-constructed English corpus of Central Asia, the average difference between the resultant values of the proposed method for calculating text similarity and Miller's human-determined values is 0.0583, and the overall deviations are all controlled within 0.15, which is basically in line with Miller's human-determined values. In terms of accuracy rate, Recall and F1 value, the proposed word weight calculation method has certain superiority. The study can identify the lexical and syntactic preferences of Central Asian English for business and diplomatic people, reduce misunderstanding, and improve the efficiency of cross-cultural communication.

Keywords: regional variation; English in Central Asia; text similarity algorithm; ITFIDF method; feature weights

1. Introduction

Looking back at the development of language policy in Central Asia, during the Soviet period, as members of the Union Republics of the USSR, the Central Asian countries could only implement the language policy set by the central government: to reduce linguistic differences among ethnic groups, to promote and consolidate the predominance of Russian as a lingua franca in the country [1-3]. This led to a movement of revival of national languages and de-Russianization in Central Asia after the collapse of the Soviet Union. The constitutions of all five countries stipulate that Kazakh, Uzbek, Kyrgyz, Turkmen, and Tajik are the national languages of their countries, and in addition to the constitutional provisions, each country has also strengthened its national language legislation [4-6]. In the 21st century, with the increasingly frequent political, economic, cultural, scientific and technological interactions among countries around the world and the globalization of English language use, the historically neglected dialects of English are gradually developing in East Asian countries and showing certain regional variations, which have become the forefront of linguistic research [7-10].

However, the previous depiction methods based on subjective feature induction cannot effectively reveal the nonlinear similarity relationship between individual variants, and the text similarity algorithm, as is an important technology in the field of natural language processing, provides support for solving the above problems. Text similarity algorithm can not only improve the effect of information retrieval and optimize the engine results, but also can be applied in the field of text

classification, machine translation, intelligent Q&A and other fields to improve the performance of related tasks [11-13]. At the same time, the research of text similarity algorithm also has an important role in promoting the development and application of natural language processing technology, in analyzing the regional variation of English dialects in Central Asia, text similarity algorithm can be effective in the five Central Asian countries English corpus, such as social media discourse, news text, etc., in the regional variant patterns [14-17].

In this paper, we adopt a cross-language text similarity calculation method based on text-weighted word co-occurrence relations, construct a cross-language word co-occurrence relation model through a parallel corpus, and then use the model for cross-language text mapping, and then calculate the similarity of texts in different languages. At the same time, a clustering algorithm based on conceptual and semantic similarity is proposed, which effectively solves the problem of "cross-cultural expression differences" and facilitates the calculation of document similarity. Then, from the perspective of inter-category and intra-category discretization, TFIDF is improved to obtain ITFIDF feature weights. Finally, the validity of this paper's method is verified in the self-built Central Asian English corpus ECCA and GlwWbE corpus.

2. Text similarity calculation

2.1. General framework

In this paper, the co-occurrence mapping model of keywords is well calculated based on the parallel corpus, and the keyword correlation relations are extracted from the articles to be detected and the newly added articles are deposited into the database, so that the computational efficiency can be improved. The algorithm application process is divided into mapping phase and matching phase, the keywords are mapped to the target text, and then the relationship matrix of other languages is mapped by the cross-language mapping model, based on which the matching text to be matched is computed, and the process is shown in Fig. 1.

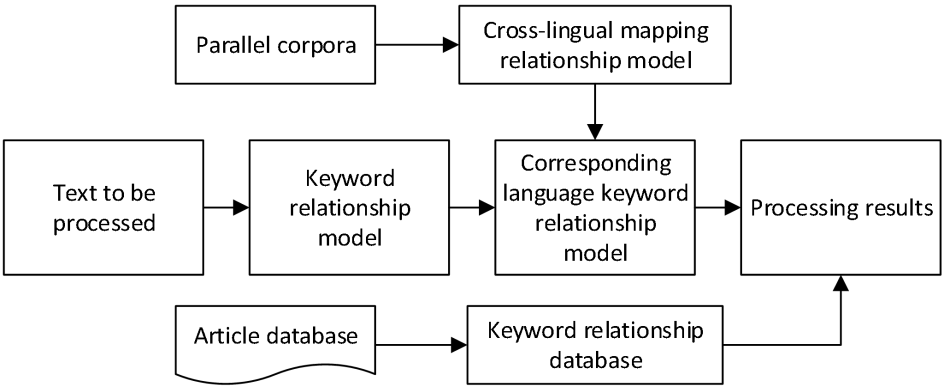

Figure 1. Algorithm flow

2.2. Mapping Relationship Modeling

This paper constructs mapping relationships across linguistic variants based on textual similarity and argues that semantic determination is based on context. The concept of context was first proposed by the British folklorist Malinowski, who believed that context is essential for understanding language. The same word may represent different meanings in different contexts, for example, Hang in English means "hang, hanging". Therefore, in this paper, we use the sentence as the smallest unit to determine the context, and use the weighted co-occurrence relationship of real words in a sentence as the basis for the construction of cross-language variant keyword mapping relationship model. Therefore, for corpus selection, a sentence-aligned parallel corpus must be chosen as follows:

(1) Count all the sentences in L_1 in which the word w_1 occurs, forming the set S_1. $S_1 = \{s_1, s_2, s_3, \cdots, s_n\}$, where s_i is a sentence containing w_1.

(2) Count the frequency of occurrence of each word in S_1. f_i, filter out the words whose frequency is too low, and sort the filtered words to obtain a new set: $F_1 = \{\langle w_1, f_1\rangle, \langle w_2, f_2\rangle, \langle w_3, f_3\rangle, \cdots, \langle w_n, f_n\rangle\}$. where:

$$f_i = \frac{\sum_{w_i \in s_i} s_i}{\sum_{i=1}^{n} s_i} \tag{1}$$

(3) Find all the sentences in L_2 that correspond to S_1, forming the sets S_1', $S_1' = \{s_1', s_2', s_3', \cdots, s_n'\}$.

(4) Do the same as in step (2) for S_1' to obtain F_1'. $F_1' = \{\langle w_1', f_1'\rangle, \langle w_2', f_2'\rangle, \langle w_3', f_3'\rangle, \cdots, \langle w_n', f_n'\rangle\}$.

(5) Save the $\langle F_1, F_1'\rangle$ mapping relation generated from the results of steps (2) and (4).

(6) Perform step (1) to (5) for all the words in L_1 to generate a L_1-to-L_2 mapping model.

Where L_1 and L_2 represent different two languages, and S_1 and S_1' represent different language-aligned sentences in L_1 and L_2, respectively.

After the processing of the parallel corpus as described in the above procedure is completed, the cross-language mapping relation model of L_1 to L_2 is obtained. If the mapping relation model of L_2 to L_1 is needed, the same process is performed for L_2.

The model actually reflects the probability distribution of keywords mapped to another language when certain keywords co-occur in one language, and can effectively solve the problem of whether to choose “AB”, “AC” or ‘BC’ as the co-occurring word pairs to be mapped when “ABC” occurs in a sentence in the dual-keyword co-occurrence algorithm, “AC” or “BC” as the co-occurring word pairs in the dual-keyword co-occurrence algorithm. The cross-language text similarity calculation is realized based on the cross-language mapping relationship model proposed in this paper.

2.3. Calculation process

The similarity calculation used in this paper is based on the cross-language variant mapping relationship model constructed in the previous section. Different from the traditional text similarity calculation method, before using this paper's algorithm for calculation, the document database to be retrieved should be preprocessed, and each document is represented by the frequency of keyword distribution to form a retrieval match vector, as follows:

(1) Sentence splitting of the L_1 language T_1 to be retrieved and splitting of T_1 into the form of the sentence set representation, i.e., $T_1 = \{s_1, s_2, s_3, \cdots, s_n\}$.

(2) Co-occurring word content and frequency are counted on a sentence-by-sentence basis for each word in T after deactivation.

$$f_{\langle w_x, w_y\rangle} = \frac{\sum_{\langle w_x, w_y\rangle \in s_i} s_i}{\sum_{i=1}^{n} s_i} \tag{2}$$

Get $F = \{\langle\langle w_1, w_2\rangle, f_{\langle w_1, w_2\rangle}\rangle, \cdots, \langle\langle w_m, w_n\rangle, f_{\langle w_m, w_n\rangle}\rangle\}$.

(3) Set the frequency threshold θ and filter out the co-occurring word pairs of $f_{\langle w_x, w_y\rangle} < \theta$, counting as vector N, where the length of N is n.

(4) For each co-occurring word pair in step (3), map it into a vector corresponding to language L_2 according to the cross-language relational mapping model, and intercept the top n results, and combine all vectors into matrix M.

(5) Compute the matrix product result $N \cdot M^T$, where M^T is the transpose matrix of M.

(6) Combine the keywords with the same frequency of the product result, count all $\langle$Keywords, frequency$\rangle$, count as $\langle r, f\rangle$ and sort the results from largest to smallest in terms of frequency to get the corresponding L_2 language co-occurring word distribution probability vector R of T.

(7) Calculate the co-occurring word distribution probability R' for each article in the database, calculate the Euclidean distance d between R and each article R', and sort the results from largest to smallest to be the similarity calculation results. Among them:

$$d=\sqrt{\left(f_{w_1}-f'_{w_1}\right)^2+\cdots+\left(f_{w_n}-f'_{w_n}\right)^2} \tag{3}$$

The central idea of the above calculation process is to map the text T of L_1 language into the probability of co-occurring word distribution of L_2 languages according to the keyword co-occurrence mapping model, and then complete the similarity calculation among documents by calculating the similarity degree of co-occurring word distribution probability of each text of L_2.

The text library to be retrieved in L_2 language can be used for co-occurring word distribution calculation, and all documents are represented by co-occurring word distribution probabilities and stored in another co-occurring word database. When retrieval is performed, the data can be obtained directly from this database, thus improving the computational efficiency. For the new text library, it can also be directly aligned with the co-occurrence probability representation, and deposited into the two databases at the same time.

3. Clustering feature extraction across linguistic variants

3.1. Feature clustering

Clustering algorithms are usually categorized into two types: hierarchical and divisional methods. The hierarchical method can be realized in two ways, merging and dividing. In merging, each sample is first made into a class, and then the number of categories is reduced by merging different classes until the termination condition is met. In split, all samples are first grouped into one class, and then the number of classes is increased by subsequent splits until the termination condition is met. Commonly used clustering algorithms are IST (IntraCluster Similarity Tech-nique), CST (Centroid Similarity Technique), UPGMA.The division method is to assign the data into a fixed number of non-empty clusters, all at the same level, the most typical division method is K-mean The most typical division method is K-mean clustering and its various variations.K-mean clustering is to randomly remove c samples from the sample as the initial cluster centers.In each iteration step, the center of mass of each class is computed, and each sample is grouped to the nearest center of mass until there is no more change in the clusters.

The text similarity matrix is a symmetric matrix and has non-zero similarity. Before clustering, the similarity matrix represents a connectivity graph. The algorithm uses the idea of hierarchical clustering with splitting to first group all the texts into a cluster, and then in each iteration, the matrix elements that do not satisfy the threshold are set a flag indicating that these two nodes are no longer connected to each other. Reconstruct the connected components of the matrix, if the number of connected components is greater than or equal to the number of input matrices K, the loop stops, otherwise the splitting continues. When the splitting stops, in order to solve the problem that the non-adjacent nodes in a cluster are not necessarily similar, a complete graph containing the maximum number of nodes is found in the corresponding connectivity graph of each cluster, which ensures that the individual nodes in the cluster must be similar. Finally, the similarity of each non-cluster node to each cluster is calculated and it is categorized into the cluster with which it is most similar. The similarity between a non-cluster node and a cluster is defined as follows:

$$Sim(d,C)=\frac{1}{|C|}\sum_{d_i\in C}Sim(d,d_i) \tag{4}$$

where d is a non-clustered node, C is a cluster and $|C|$ denotes the number of nodes in the cluster.

3.2. Feature weight calculation

The common method of feature word weight is word frequency-inverse document frequency, which well reflects the degree of a word's contribution to a specific document, and is widely used as an effective weight calculation method. Therefore, the feature topic weights refer to the feature word weight calculation method, and the Topic Frequency-Inverse Document Frequency (TFIDF) is used to calculate the feature topic weights.

3.2.1. TFIDF

The TFIDF is calculated as:

$$TFIDF = TF \times IDF = P\left(Z_k \left| d^i \right.\right) \times \log \frac{\left|M\right|}{1+\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|} \tag{5}$$

where TF is the topic frequency, i.e., how often topic Z_k appears in document d^i, i.e., $P\left(Z_k \left| d^i \right.\right)$; and IDF is the inverse document frequency, i.e., $\log \frac{\left|M\right|}{1+\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|}$. $\left|M\right|$ represents the total number of documents in the document set, $\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|$ represents the number of documents containing topic Z, λ is a set constant, and $P\left(Z_k \left| d^i \right.\right) = \theta_{d^i,k}$, the formula reflects the extent to which the topic contributes to a particular document.

According to the definition of text similarity, similarity documents must belong to the same category, so the above formula has some shortcomings, its intra-category and inter-category differentiation ability is not strong, can not be given to represent the category of the characteristics of the theme of high weight, resulting in the same category of similarity documents match weaker. For example, category C contains more documents with theme Z, and other categories contain less, indicating that theme Z can represent category C, that should be given a higher weight. And formula (5) does not increase the weight of theme Z that can represent category C, because when the C category theme Z increases, $\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|$ will also increase, and the value of IDF will decrease, at the same time, when the feature theme is uniformly distributed in category C should also be given a high weight, so on the basis of this, it is improved $TFIDF$.

3.2.2. Improvement of TFIDF

The TFIDF is improved to obtain the ITFIDF from the perspective of inter-class and intra-class discretization.

(1) The traditional IDF does not consider the distribution of feature topics between classes, and the IDF is first improved by increasing the weights of feature topics that appear more frequently in a class. The improved IDF is:

$$IDF = \log \frac{\left(\left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right| + 1\right) \times \left|M\right|}{1+\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|} \tag{6}$$

where $\left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|$ is the number of documents containing topic Z in category C, $\left|C\right|$ is the number of all documents in category C, set $\left|d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right| = \left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right| + \left|O\right|$, $\left|O\right|$ represents the number of documents containing topic Z in other categories, set:

$$f\left(\left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|\right) = \frac{\left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right| + 1}{1+\left|C^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right| + \left|O\right|} \tag{7}$$

Then as $\left|C_d^i : P\left(Z_k \left| d^i \right.\right) > \lambda\right|$ increases, the f value increases in tandem, i.e., raising the IDF value. Equation (7) then takes into account the distribution of topic Z between classes, increasing the degree of contribution to the category to make it better represent the category of documents, and adding 1 to prevent the numerator from appearing 0.

(2) Also, IDF does not consider the distribution of clustered features within a class. Therefore, this paper analyzes the distribution of feature themes within a class by examining the degree of dispersion of the clustered features within the class D. The degree of dispersion can reflect the distribution of clustered features in a class, which is expressed by the standard deviation of the probability distribution of the themes, and is improved as follows:

$$\overline{P}'\left(Z_k \left| d^i \right.\right) = \frac{1}{\left|C\right|} \sum_{i=1}^{|C|} P\left(Z_k \left| d^i \right.\right) \tag{8}$$

$$D=\sqrt{\frac{1}{|C|}\sum_{i=1}^{|C|}\left(P\left(Z_k\left|d^i\right.\right)-\overline{P}'\left(Z_k\left|d^i\right.\right)\right)^2} \tag{9}$$

Equation (8) represents the mean value of the probability distribution of topic Z in category C documents, and equation (9) represents the degree of dispersion of the distribution of topic Z in category C documents. A lower value of formula (9) means that theme Z is more evenly distributed in category C, which means that theme Z is more representative of category C. Since the degree of intra-class dispersion is inversely proportional to the categorization ability of the featured topics, the intra-class correction formula is replaced by $(1-D)$, and the final $ITFIDF$ is:

$$ITFIDF=P\left(Z_k\left|d^i\right.\right)\log\frac{\left(\left|C_d^i:P\left(Z_k\left|d^i\right.\right)>\lambda\right|+1\right)\times|M|}{1+\left|d^i:P\left(Z_k\left|d^i\right.\right)>\lambda\right|}(1-D) \tag{10}$$

With the improvement, Eq. (10) then enhances the ability to differentiate between and within classes, resulting in an increase in the weight of topics representing a particular class.

4. Experimentation and analysis

In order to verify the effectiveness of the methodology of this paper, two groups of comparable corpora were selected for experimental validation. The first group is the self-constructed English Corpus of Central Asia (ECCA), which covers the official English pages of universities, government press releases in English and English abstracts of academic papers in five Central Asian countries, including Kazakhstan and Uzbekistan, totaling about one million words. The second group is a subset of British English (UK) and a subset of South Asian English (India, Pakistan) extracted from the GlwWbE corpus, which represent Standard English and English variants that are geographically close and have similar language contact environments, respectively. In this paper, the improved weight calculation method ITFIDF is used to select the feature words, and the traditional text similarity algorithm as well as the algorithm combining information gain and intra-class discretization, which is mainly improved in this paper, are used to verify the accuracy of various algorithms through experiments respectively. Experimental environment: CPU Intel Core 3.40Ghz, memory 16 G, operating system Windows10 64-bit, development environment Python 3.5, word separation tool THULAC.

4.1. Similarity calculation results

In order to verify the feasibility of the text similarity method proposed in this paper, this paper selects 10 pairs of English words in two corpora, ECCA and GlwWbE, and uses this paper's method to calculate the semantic similarity of words. Firstly, these 10 pairs of English words are translated into corresponding Chinese word pairs according to the principles of homonymy and closest meaning, and then the similarity between the two groups of words is calculated by the method of this paper, and finally compared with the value of Miller's manual judgment. The range of semantic similarity is determined as [0,1], with 0 indicating that the words are completely different in semantics and 1 indicating that they are identical, and the comparison of the calculated results is shown in Figure 1.

Table 1. Results of Word Similarity Calculation

Number	Words		Text similarity	Similarity range	Miller's judgment	Difference
	ECCA	**GlwWbE**	**0.5101**	**0.5-0.6**	**0.6161**	**0.1060**
1	Flat	Apartment	0.8836	0.8-0.9	0.8260	0.0576
2	autumn	fall	0.8064	0.8-0.9	0.8577	0.0513
3	lift	elevator	0.8196	0.8-0.9	0.8196	0.0040
4	petrol	gasoline/gas	0.7138	0.7-0.8	0.7039	0.0099
5	biscuit	cookie	0.5455	0.5-0.6	0.5056	0.0399
6	crisps	chips	0.7028	0.7-0.8	0.6088	0.0940
7	lorry	truck	0.5476	0.5-0.6	0.5938	0.0462
8	torch	flashlight	0.5166	0.5-0.6	0.6107	0.0941
9	queue	line	0.7658	0.7-0.8	0.8020	0.0362
10	rubbish	garbage/trash	0.6659	0.6-0.7	0.5634	0.1025

As can be seen from the table, the average difference between the result value of text similarity calculated by the method proposed in this paper and the value determined by Miller is 0.0583, indicating that the value calculated by the text similarity calculation method proposed in this paper is

basically in line with the value determined by Miller manually, and only a few of them are due to the existence of a certain degree of error in the translation, which leads to a higher relative bias of the results, and the overall bias is controlled at 0.15 The overall deviation is controlled within 0.15. The data results show that the results of text similarity calculation using the cross-language variant mapping relationship model have high accuracy and reflect human subjective judgment, which verifies that it is feasible to calculate the similarity between cross-language variant words using this method, and it can provide an effective method and way to measure the semantic similarity of words.

4.2. Analysis of threshold clustering results

The experimental data in this paper comes from the ECCA and GlwWbE corpora, from which three clustering features are selected, covering 100 articles each in three categories: official university English (Type I), government English press releases (Type II) and academic English (Type III). The number of clusters k in the experiment is taken as 6, and the ratio of feature value selection Q in the clustering process in chapter 3.1 is taken as 50%. In order to ensure that the experiment achieves the best results, the value of the threshold μ needs to be determined. This experiment is divided into six groups, in respectively take the threshold p for 0.50, 0.55, 0.60, 0.65, 0.70 and 0.75 when the relevant text clustering experiments, respectively, each group threshold p under the calculation of the corresponding F value, each group of experiments for five times the cross-experiment, the best clustering effect of the selected one as the final result.

Figure 2 gives the experimental results of the effect of different thresholds μ on the effect of the same clustering algorithm. At first, as the threshold p of text similarity between keyword items gradually increases, the F-value for judging the clustering effect gradually rises, reflecting that the clustering effect is getting better and better. Until the threshold μ reaches a critical region, that is, the threshold μ is at about 0.60, the F-value reaches the highest, at this time, the feature clustering effect is the best, but continue to increase the threshold μ, the feature clustering effect instead of becoming worse. This is because the similarity calculation value of the keyword items is rarely more than 0.65, which affects the weighting effect of the keyword items, thus affecting the results of the similarity calculation, and it can be seen from the figure that the F-value decreases rapidly after the threshold p exceeds 0.65.

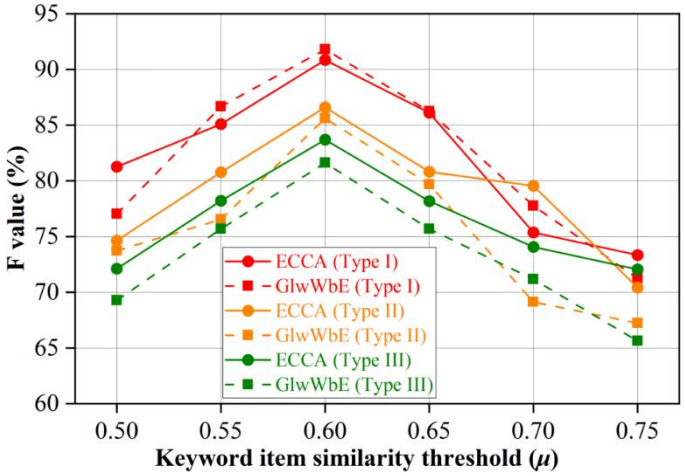

Figure 2. The influence of similarity threshold μ on the effect of feature clustering

4.3. Analysis of feature weighting results

In order to verify the feature weight calculation method based on ITFIDF proposed in this paper, this paper combines the TFIDF method with information gain (IG-TFIDF), the TFIDF method with intra-class discretization (DI-TFIDR), and the improved weight calculation method in this paper to select the features, and compare them experimentally through the KNN classification method, and the experimental results are shown in Table 2. shown.

Table 2. Calculation Results of Feature Weights

N	Precision	Recall	F1-score

	IG-TFIDF	DI-TFIDR	This paper	IG-TFIDF	DI-TFIDR	This paper	IG-TFIDF	DI-TFIDR	This paper
N1	0.8328	0.6959	0.7678	0.6973	**0.9242**	0.8517	0.7383	0.6819	**0.7940**
N2	0.8759	0.6336	**0.9134**	0.6574	0.7306	**0.8192**	0.6523	**0.8450**	0.6084
N3	0.6263	0.6598	**0.6948**	**0.9206**	0.6152	0.5680	0.7305	**0.8979**	0.7078
N4	0.6333	0.8803	**0.9715**	0.6409	0.6363	**0.9526**	**0.7376**	0.6083	0.7252
N5	0.8639	**0.9571**	0.6117	0.6345	0.8422	**0.8810**	0.7477	0.8347	**0.9845**
N6	**0.8986**	0.8363	0.8243	0.8292	0.6011	**0.8966**	0.8834	0.7678	**0.9181**
N7	0.6006	0.5917	**0.7008**	**0.9412**	0.9399	0.8887	0.6378	**0.8019**	0.7588
N8	0.6020	0.7571	**0.8154**	0.5858	**0.8521**	0.8394	0.6633	0.7623	**0.8267**
N9	0.8373	0.5620	**0.8502**	0.7378	0.6176	**0.8599**	**0.9320**	0.6297	0.8560
N10	0.6307	0.6868	**0.9072**	0.8756	**0.9808**	0.6348	**0.9897**	0.6155	0.8836
Mean	0.7401	0.7261	**0.8057**	0.7520	0.7740	**0.8192**	0.7713	0.7445	**0.8063**

According to the table, it can be seen that in terms of precision rate, the word weight calculation method proposed in this paper has achieved the maximum value in the seven categories of N2, N3, N4, N7, N8, N9 and N10, with an average precision rate of 0.8057, which is also more than the 0.7401 of the IG-TFIDF and the 0.7261 of the DI-TFIDR.In terms of Recall and F1 values, the weight calculation method proposed in this paper also exceeds IG-TFIDF and DI-TFIDR by 0.7261 in terms of the average value, which shows that the word weight calculation method proposed in this paper has a certain degree of superiority.

The above results are only evaluated for the classification effect of a specific class, so the classification effect of the whole text category will be evaluated by Macro-P, Macro-R, Macro-F1 and Accuracy, and the results are shown in Table 3.

Table 3. Overall Accuracy, Recall Rate, F1 Value (%)

	IG-TFIDF/%	DI-TFIDR/%	This paper/%
Macro-P	84.03	85.96	**88.89**
Macro-R	83.28	85.33	**87.52**
Macro-F1	83.35	85.46	**87.39**
Accuracy	84.31	86.28	**87.61**

From the table, it can be seen that using the feature weight calculation method proposed in this paper, compared with using the IG-TFIDF, and DI-TFIDF methods, in terms of Macro-P, it improves 4.86% and 2.93%; in terms of Macro-R, it improves 4.24% and 2.19%; in terms of Macro-F1, it improves 4.04% and 1.93%; in Accurary, 3.3% and 1.33% respectively It can be seen that the feature weight calculation method proposed in this paper is improved on the basis of both IG-TFIDF and DI-TFIDR methods. The experiment confirms the effectiveness of the word weight calculation method proposed in this paper.

5. Conclusion

In this study, a quantitative analysis framework oriented to the regional variation of English dialects in Central Asia is constructed based on the text similarity algorithm. By participating in controlled experiments with the constructed English corpus ECCA and GlwWbE corpus in Central Asia, combined with ITFIDF feature weight calculation, the following conclusions are obtained:

(1) This paper describes in detail the process of constructing a word co-occurrence mapping model based on the corpus, as well as the process and experimental flow of text similarity calculation based

on the word co-occurrence mapping model. The experimental results show that the overall deviation of the cross-language text similarity calculation of the proposed method in this paper is controlled within 0.15, which is closer to the manual judgment standard.

(2) The algorithm only requires that the source language and target language can be split into words, and there is a certain amount of parallel text to do cross-language text similarity computation, which effectively reduces the time complexity and dimensionality, and mines the potential semantics of the document without the need for bilingual knowledge and relevant linguistic features.

(3) The study fills the gap of empirical research on the "Central Asian region" in the study of world English, and also provides computational linguistics evidence to support the expansion of intra-circle heterogeneity. In the future, we can further introduce dependency features and depth representation models, and expand the corpus of spoken dialogues to reveal the evolutionary variation of Central Asian English at the phonological and pragmatic levels more comprehensively.

Funding

Funded Project: 2022 Special Project for Enhancing Comprehensive Discipline Strength at Yili Normal University: Research on the Needs and Countermeasures for Emergency Language Services in Xinjiang under the Belt and Road Initiative (Project Number: 22XKSY40).

References

1. Nogayeva, A. (2015). Limitations of Chinese "Soft power" in its population and language policies in Central Asia. Geopolitics, 20(3), 583-605.
2. Reagan, T. (2019). Language planning and language policy in Kazakhstan. In The Routledge international handbook of language education policy in Asia (pp. 442-451). Routledge.
3. Bahry, S. (2019). Towards "mapping" a complex language ecology: The case of Central Asia. In Handbook of the changing world language map (pp. 3-41). Cham: Springer International Publishing.
4. ZHILTSOV, S., SLISOVSKIY, D., SHULENINA, N., & MARKOVA, E. (2018). Shaping national Identities in Central Asian countries: Results, Problems, Prospects. Central Asia & the Caucasus (14046091), 19(1).
5. Bahry, S., Niyozov, S., Shamatov, D. A., Ahn, E., & Smagulova, J. (2017). Bilingual education in Central Asia. Bilingual and multilingual education, 259-280.
6. Kirkpatrick, A., & Liddicoat, A. J. (2019). Language education policy in Asia: An overview. The Routledge international handbook of language education policy in Asia, 3-13.
7. Bolander, B., & Sultana, S. (2019). Ordinary English amongst Muslim communities in South and Central Asia. International Journal of Multilingualism, 16(2), 162-174.
8. Bahry, S. A. (2016). Language Ecology: Understanding Central Asian Multilingualism. Language change in central Asia, 106(11).
9. Jehan, N., Aurangzeb, S., & Javid, T. (2025). Linguistics Variations in Teaching Learning Process: An Analytical View of Education System in Central Asian Republics. International Research Journal of Arts, Humanities and Social Sciences, 2(02), 141-151.
10. Fierman, W. (2016). Trends of language use in prestige domainsin Post-Soviet Central Asia. Alatoo Academic Studies, (1), 35-42.
11. Prasetya, D. D., Wibawa, A. P., & Hirashima, T. (2018). The performance of text similarity algorithms. International Journal of Advances in Intelligent Informatics, 4(1), 63-69.
12. Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. Machine Learning and Applications: An International Journal, 3(2), 19-28.
13. Liu, Y., & Chen, M. (2021). Applying text similarity algorithm to analyze the triangular citation behavior of scientists. Applied Soft Computing, 107, 107362.

14. Lewis, M., Cahill, A., Madnani, N., & Evans, J. (2023). Local similarity and global variability characterize the semantic space of human languages. Proceedings of the National Academy of Sciences, 120(51), e2300986120.
15. Wang, J., & Dong, Y. (2020). Measurement of text similarity: a survey. Information, 11(9), 421.
16. Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications, 120(9), 29-34.
17. Biloshchytska, S., Tleubayeva, A., Kuchanskyi, O., Biloshchytskyi, A., Andrashko, Y., Toxanov, S., ... & Sharipova, S. (2025). Text similarity detection in agglutinative languages: A case study of Kazakh using hybrid n-gram and semantic models. Applied Sciences, 15(12), 6707.