axioms

MDPI

Article

The Performance of Topic Evolution Based on a Feature Maximization Measurement for the Linguistics Domain

Junchao Feng [1,2,*], Jianjun Miao [1], Yue Tang [3], Yuechen Li [4] and Jundong Feng [1]

[1] School of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China
[2] Foreign Language Department, Harbin University of Science and Technology, Harbin 150080, China
[3] Alibaba Beijing Software Co., Ltd., Beijing 100020, China
[4] JD Group, Beijing 100176, China
* Correspondence: fjc2021@nuaa.edu.cn

Abstract: Understanding the performance of the data mining approach and topic evolution in a certain scientific domain is imperative to capturing key domain developments and facilitating knowledge transfer within and across domains. Our research selects linguistics as an exploratory domain and exploits the feature maximization (FM) measurement for feature selection, combined with the contrast ratio to conduct the diachronic analysis for the linguistics domain's topics. To accurately mine the linguistics domain's topics and obtain the optimal clustering model selection, we exploit an integrated method associated with the deep embedding for clustering (DEC) algorithm based on the keywords-based Text Representation Matrix (KTRM) and Lamirel's *EC* index and test the performance of this method. The results show that the FM measurement is applicable in the linguistics domain for topic mining, and the combinatory method has the advantage of an unbiased clustering optimization model and applies to the design of non-parameter clustering and algorithms from the low dimension to the high dimension of datasets. The findings suggest that this approach could be suitable for a diachronic analysis of topic evolution and facilitate the performance of topic detection. In addition, these findings of text detection can rise to knowledge fusion cognition with the factor of language as an available research objective in interdisciplinary research.

Keywords: feature maximization; *EC* index; deep embedding clustering (DEC); linguistics domain; topic evolution

MSC: 03C98; 00A69; 00A71

Citation: Feng, J.; Miao, J.; Tang, Y.; Li, Y.; Feng, J. The Performance of Topic Evolution Based on a Feature Maximization Measurement for the Linguistics Domain. *Axioms* **2022**, *11*, 412. https://doi.org/10.3390/axioms11080412

Academic Editor: Sidney A. Morris

Received: 6 June 2022
Accepted: 25 July 2022
Published: 18 August 2022

1. Introduction

Nowadays, with the fast speed of knowledge transmission, how to quickly master a panorama of diverse scientific domains has attracted considerable attention in the academic world. Topic analysis in diverse scientific domains is important to clarify and identify emerging topics, hot topics, and knowledge transfer [1]. Topic analysis combined with an effective approach could provide an appropriate research strategy for various research purposes, such as the convergence and divergence of research themes [2], identifying experts [3], exploring interdisciplinary topics [4–7], detecting research events [8,9], and community detection [10]. Aside from that, topic evolution analysis is conductive to thoroughly understanding the diachronic change and predict paradigm shifts in disciplinary development [11]. Specifically, it can do a favor for researchers in discovering academic topics through extracting and summarizing trending topics combined with effective topic detection approaches in the form of useful information and help track topic-based communities, further optimize research topic choices, and seek scientific collaborators [10]. Moreover, it can facilitate the promotion of knowledge transmission within or across diverse domains [11].

Previous studies suggest that a growing number of studies have used various topic detection approaches to give rise to topic analysis and topic evolution analysis. Academic research on topic detection methods aims at helping to analyze a set of documents with a relatively formulated frame and balanced data. However, such usual approaches are less suitable for documents with sparse data or unexpected noise because whenever the dataset is constituted by complex data which need to be represented in both a high-dimensional and sparse description space, it is difficult to identify an optimal clustering model [12,13]. That aside, some extant studies only focus on topic evolution analysis for certain nature science domains rather than that of humanities and social sciences by an optimal topic model based on a feature maximization index [14–17].

Linguistics, as one of the important subjects of the humanities and social sciences domains, is a research domain related to language and language use which is highly complex and cross-disciplinary. It enables human beings to create or expand into new domains. Thus far, although the academic circle has used data-driven paradigms to make exciting research topic discoveries in many fields [18,19], most research on the linguistics domain are synchronic ones which depict topics statically, and research that investigates the topic evolution of linguistics is relatively rare and confined to a few methods, such as the introspection method and hypothesis-driven research method [20–22]. Accordingly, this study is interested in attempting the application of an unsupervised categorization framework combined with feature maximization (FM) measurement [23,24] with the deep embedding clustering (DEC) model [25] to further explore the following two questions: "what is the linguistics research interested in", and "how do such topics change over time?" Our proposed method utilizing the DEC technique can increase the distance between different kinds of documents and decrease the distance between similar documents to improve the accuracy of clustering. We can achieve higher linguistics topic clustering accuracy which is better than those of general clustering algorithms such as K-means, GMM, and so on, especially when we deal with high-dimensional data. Our study cannot neglect one condition: when we analyze topic evolution in the linguistics domain, we should not only consider the optimal number of clusters but also the cluster quality index. Aside from that, the intra-class inertia and inter-class inertia of the collections of words in the same clustering are of great importance, so our study utilized an *EC* index [23], which is based on the maximization of the average weighted compromise between the contrast of active features and the inverted contrast of passive features for optimal partition to facilitate the cluster quality.

The major contributions of this study are summarized as follows. This study aims to explore "what is linguistics research interested in" and "how do such topics change over time" to advance topic analysis and topic evolution analysis in the linguistics domain, which is also based on a combinatory topic detection method of feature maximization, the contrast ratio, and the DEC clustering method based on a keyword-based Text Representation Matrix (KTRM) [26]. First, this study suggests the critical role of the combinatory approach to topic detection and topic evolution analysis in the linguistics domain. Second, this study supports the *EC* index, in which utilizing both active and passive features is proven to have better performance that is especially suitable for producing stable results and requires little computation time for processing high-dimensional text data, which provides a reference for researchers without consideration for clustering parameter estimation when analyzing the diachronic evolution of topics for a certain discipline. Third, this study explores topic analysis and topic evolution analysis in the linguistics domain, wherein the extant literature has scarce research-based knowledge on the linkage between feature maximization measurement and the performance of topic evolution in a visualized way. Lastly, this study contributes to the aims of knowledge mapping visualization based on the combinatory method, presenting a panoramic view of the topic evolution and the relative relationship of each clustering topic in the exploratory study of the linguistics domain.

The rest of the article is organized as follows. Section 2 provides the related work on topic detection methods and studies on topic evolution. Section 3 highlights the selected

data and the methodology for data discovery, as well as extraction of the F-value. Next, Section 4 conducts topic detection and contrast graph visualization of linguistics research based on this combinatory approach. Section 5 follows discussions of the results according to the research scheme above. Section 6 provides the conclusion and limitations of the study.

2. The Related Work

There have been many approaches from a macro level to a micro level to describe topic evolution in analyzing domain knowledge. Researchers in diverse domains are interested in varied topics over time, resulting in topic evolution [27,28]. Mane and Börner [29] employed a topic detection and tracking method to find topic dynamics in document sequences. Blei and Lafferty [30] studied the top topics of science in scientific papers. They provided a macro-level picture of topic evolution. Aside from that, Chen et al. [27] used an analogy to detect the evolution of topic splitting or topic merging on a detailed level. Generally, through the approaches of topic detection and tracking analysis, researchers would introduce a knowledge domain visualization technique to extensively identify and map for holistic domain knowledge mapping from the scientific literature. However, to accurately detect the domain research topics, traditional clustering methods have certain limitations and have difficulty dealing with complex high-dimensional data. Therefore, deep learning clustering methods have begun to attract extensive attention from researchers.

For text data, the effectiveness of current clustering methods largely depends on the quality of input text representation; that is, text features are transformed into quantified representations that can be recognized by computers. If the dataset has a growing size, it is difficult to find the more accurate ground truth, which works on them in supervised clustering. To avoid this issue and increase the quality of the clustering, this study exploits the FM measurement for feature selection combined with the contrast ratio to conduct diachronic analysis for the linguistics domain's topics.

2.1. Feature Maximization Combined with the Contrast Ratio for the Selected Feature

Studies on topic evolution in the extant research articles derived from the discovering topic detection approach are presented. Content analysis has been shown to be capable of classifying these data from published articles [31]. However, with the massive publications available today, topics can no longer be detected or summarized by human annotation. Therefore, sophisticated approaches assisted by clustering or topic modeling algorithms appeared to facilitate topics' extraction. According to the pioneering contribution of Lamirel's FM combined with the contrast ratio, Lamirel et al. [23] also proposed the *EC* quality index to test data from a multisource bibliographic database in order to deal efficiently with diachronic analysis, and the findings confirmed the clear advantages of the *EC* quality index, which could help to find stable results in cases ranging from low-dimensional to high-dimensional contexts and also save enough computation time while easily dealing with binarized data. In addition, Chen et al. [18] directly adopted the FM measurement combined with GNG clustering to analyze the topic evolution of 40 years of science in China. Their experiments also confirmed that the FM measurement is helpful for feature selection and the unsupervised clustering quality.

Specifically, FM, which was initially proposed by Lamirel et al. [32], is an unbiased cluster quality metric that is used to exploit the data features associated with each cluster. It has been proven that its significant advantage in the process of clustering is to be independent of the clustering method and its operating model [33]. Additionally, Lamirel et al., challenged the optimal model selection by relying on feature maximization, which can provide a highly efficient feature selection and feature contrasting model [33] to estimate the quality of classification without prior consideration of the cluster profiles. Interestingly, their research can obtain this in the case of the classification of highly unbalanced, highly multidimensional, and noisy data [34]. Therefore, FM is a cluster quality metric that is in favor of clusters with maximum feature representation of their associated data.

Because the main advantage of the FM is to be independent of the clustering method and its operating model [34], consequently, this article combines unsupervised deep embedding for clustering (DEC) with the feature maximization algorithm to extract the features represented by the nouns from the title, abstract, and keywords of linguistics bibliographic retrieval and cluster them into diverse categories. Consider a partition C which results from a clustering method applied to a database D, represented by a group of features F. The feature maximization measure favors clusters with a maximal feature F measure. The feature F measure $FF_c(f)$ is the harmonic mean of the feature recall $FR_c(f)$ and the feature predominance $FP_c(f)$, which can be thought of as follows:

$$FR_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{c \in C} \sum_{d \in c} W_d^f} \quad FP_c(f) = \frac{\sum_{d \in c} W_d^f}{\sum_{f' \in F, d \in c} W_d^{f'}} \tag{1}$$

with

$$FF_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right) \tag{2}$$

where the harmonic mean here can provide an additional influence on the lowest of the two values in the combination of feature recall and feature precision, according to Dempster et al. [35]. That aside, W_d^f stands for the weight of the feature f for element d, and F_c represents the set of features associated with the data occurring in cluster c. $FP_c(f)$ measures the ability of f to describe cluster c. In a complementary way, $FR_c(f)$ allows characterizing f according to its ability to discriminate c from other clusters. Finally, the feature F measure $FF_c(f)$ of a cluster $c \in C$ is the average of the feature F measures of the maximal feature for c. According to exhaustive experiments of large reference datasets of bibliographic records [2], the feature maximization approach is treated as a reliable approach to solving complex high-dimensional classification problems with highly unbalanced and noisy data gathered in similar classes due to its efficient feature selection and data resampling capabilities.

Taking a parameter-free, class-based process for the feature F measure into consideration, this paper exploits a class feature that is characterized using both its capacity to discriminate a given class from others ($FP_c(f)$ index) and its capacity to precisely represent the class data ($FR_c(f)$ index) in the selection process. The set S_c of features that are characteristic of a given class c belongs to an integral class set C, which results in

$$S_c = \left\{ f \in F_c | FF_c(f) > \overline{FF}(f), FF_c(f) > \overline{FF}_D \right\} \tag{3}$$

where

$$\overline{FF}(f) = \sum_{c' \in C} \frac{FF_{c'(f)}}{\left|C_{/f}\right|} \text{ and } \overline{FF}_D = \sum_{f \in F} \frac{\overline{FF}(f)}{|F|} \tag{4}$$

in which $C_{/f}$ represents the subset C to the classes in which the feature f is represented. Subsequently, the set of all the selected features S_C is the subset of F, which is defined as

$$S_C = \cup_{c \in C} S_c \tag{5}$$

Namely, in terms of the feature F measurement, the features judged relevant for a given cluster are those whose representations are not only better than their average description in this cluster but also better than the average representation of all the features in the partition. In the specific framework of the feature maximization process, a particular contrast concept $G_c(f)$ is defined to calculate the performance of a retained feature f for a given cluster c. This contrast enhancement step can be exploited as complementary to the former feature selection step. $G_c(f)$ is an indicator value. It is proportional to the ratio between the feature F measure $FF_c(f)$ of a feature in cluster c and the average feature F measure $\overline{FF}(f)$ of this feature for the whole partition. However, it is worth noting that this measurement

would introduce unexpected Gaussian smoothing in the process. The contrast $G_c(f)$ can be expressed as

$$G_c(f) = \frac{FF_c(f)}{\overline{FF}(f)} \tag{6}$$

The research regulates that the active features of a cluster are those for which the contrast is greater than one. In addition, the higher the contrast of a feature for one cluster, the better its performance in describing the cluster content. Clarifying the rationale of the algorithms above, we illustrate the calculation method with an example based on a classic Iris dataset. There are two categories of Iris—Iris-Setosa (S) and Iris-Versicolor (V)—in the dataset. The following four features of each flower are measured: the Calyx_Length (A), Calyx_Width (B), Petal_Length (C) and Petal_Width (D). The feature F measure corresponding to each feature in each class is calculated, and then the average value of all features is obtained.

Figure 1 shows the source data and how the feature F measure calculation of the Calyx_Length (A) feature operates in the Iris-Setosa (S) class. As shown in Figure 2, the next step is comprised of calculating the average F measure of each feature over the clusters and the overall average F measure for the combination of all features and all classes. In Figure 3, $\overline{F(.,.)}$ represents the overall average $\overline{FF_D}$ presented in Equation (3), and $\overline{F(x,.)}$ stands for the average class x, which is itself computed as

$$G_c(f) = \frac{FF_c(f)}{\overline{FF}(f)} \tag{7}$$

Calyx_ Length	Calyx_ Width	Petal_ Length	Petal_ Width	Class
5.1	3.4	1.4	0.2	S
4.9	3	1.4	0.2	S
4.7	3.2	1.3	0.2	S
7	3.2	4.7	1.4	V
6.4	3.2	4.5	1.5	V
6.9	3.1	4.9	1.5	V

$$FR(A,S) = 14.7/35 = 0.42$$
$$FP(A,S) = 14.7/29 = 0.51$$
$$FF(A,S) = \frac{2(FR(A,S) \times FP(A,S))}{FR(A,S) + FP(A,S)} = 0.46$$

Figure 1. Principle of feature F measure computation for sample data.

	$F(x,S)$	$F(x,V)$	$\overline{F(x,.)}$
Calyx_ Length	0.46	0.49	0.48
Calyx_ Width	0.4	0.28	0.34
Petal_ Length	0.17	0.42	0.3
Petal_ Width	0.04	0.17	0.11

$\overline{F(.,.)}$
0.31

Figure 2. Principle of computation of overall feature F measure average and elimination of irrelevant features.

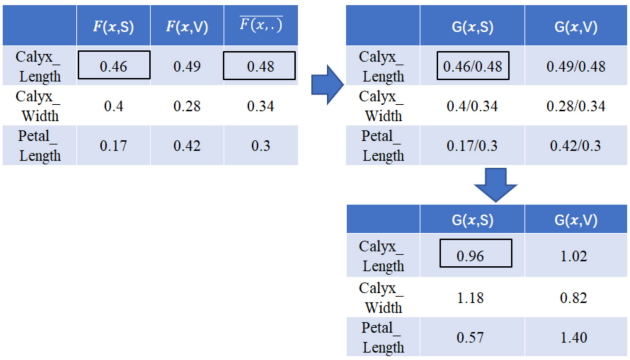


Figure 3. Principle of computation of contrast for selected features.

After calculation, when there are features with F measures that are systematically lower than the overall average, they are eliminated. Specifically, the Petal_Width feature is thus removed. The rest of the features (i.e., the selected features) are reckoned to be active features in the classes in which their F measures are above the marginal average:

(1) Calyx_Length and Calyx_width are active in Iris-Sentosa's class (S);
(2) Calyx_Length and Petal_Length are active in Iris-Versicolor's class (V).

This contrast sheds light on the degree of activity and passivity of the selected features concerning their F measure marginal average in different classes. Figure 3 illustrates how the contrast is calculated for the example above. Under the circumstance of this classic example, the contrast may be regarded as a function that will, in essence, have the following influences:

(1) A virtual increase in the width of Iris-Sentosa's calyx;
(2) An increase in the length of Iris-Versicolor's calyx and petals;
(3) Conversely, a decrease in the length of Iris-Sentosa's calyx and petals;
(4) A decrease in the width of Iris-Versicolor's calyx.

The active features in a cluster are selected for which the contrast is greater than one in that cluster, as opposed to the passive features in a cluster with a contrast less than the unity value. According to the contrast, the Clayx_Width is selected as the active feature for Iris-Setosa, in contrast to the Calyx_Length and the Petal_Length, which are the active features for Iris-Versicolor.

This method of exploiting the features obtained is employed to utilize the actively selected features and their corresponding contrast for cluster labeling [32], as Lamirel et al. proposed. Later, a more complicated method related to the activity and passivity of the selected features in clusters is used to exploit and satisfy the clustering quality indexes which can identify an optimal partition.

2.2. Deep Embedded Clustering (DEC) Combined with a Keyword-Based Text Representation Matrix (KTRM)

As is known to us all, traditional text representation methods usually exist with high data dimensionality, large sparseness, and inadequate representation of structural and semantic information, which leads to a higher time complexity and computation complexity for the text clustering. When we accurately detect the diachronic analysis of high-dimensional massive data for a certain domain, the traditional methods have outstanding limitations. Considering that the FM measurement, as mentioned above, is independent of clustering methods, our study exploits the Keyword-Based Text Representation Matrix (KTRM) [26] associated with deep embedded clustering (DEC) [25] to obtain the optimal clustering number. According to their experiment involving news text on actual agricultural product trade friction, this clustering method's clustering accuracy and standard mutual information were significantly improved. Therefore, our study aims to achieve a better clustering effect for topic evolution analysis of the linguistics domain,

where we exploit the text representation of the input data based on the KTRM and DEC for clustering.

We take inspiration from DEC (see Figure 4 for the principle of DEC), which has good robustness, universality, and migration on different datasets. This is extremely important for clustering. Then, we will enter the KTRM representation of the text based on the FM measurement combined with the contrast ratio in the interface of the image dataset processed by DEC, at which point DEC normalizes it and iteratively optimizes the feature representation and cluster allocation. However, in practice, the number of natural clusters is often unknown. Therefore, an approach to determining the optimal number of clusters is necessary. Inspired by the algorithms of Xie et al. [25] and the *EC* index proposed by Lamirel et al. [23], we would test and obtain the optimal number of clusters which is helpful for analyzing the diachronic topic evolution of the linguistics domain.

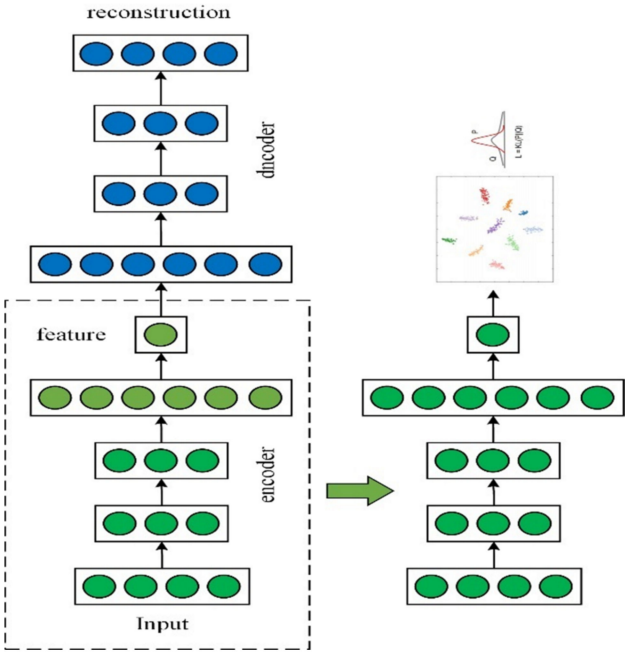


Figure 4. The principle of the DEC algorithm [25].

Consequently, our study exploits the above-mentioned methods to determine the optimal partitioning scheme, which is measured by feature maximization and specific information related to the activity and passivity of cluster features in practice. In addition, in order to lower the computational complexity and find the optimal number of clusters, we will adopt a non-parametric process in the whole feature selection process, and the clustering labels can be marked according to the most representative features. Accordingly, this study proposes two questions about linguistics research: "what is linguistics research interested in", and "how do such topics change over time?" To answer the above questions, this article will explore topic evolution in the linguistics domain based on a combinatory approach with feature maximization, the contrast ratio, and the DEC clustering method. This research uses unsupervised topic modeling techniques to detect topics derived from articles of the linguistics research in CNKI from 1999 to 2018. Additionally, the results will be shown through visualization technology with Cytoscape software, which can represent the panorama of linguistics research topics and other related interdisciplinary research that takes language as a research object.

3. Data Collection and Data Preprocessing

This study retrieved 9621 articles (core journals and CSSCI periodicals; retrieval time: 10 April 2019) derived from China's public CNKI database in the period from 1999 to 2018. Specifically, we selected "linguistics" and "language research" as the search term

to retrieve conducted data cleaning (delete articles such as those called "notification", "meeting notice", "magazine profile", "to inform the reader", and other types of literature) and refined 6639 academic papers as the research object of this article. Then, we gained the trend of the number of linguistic journal papers in Figure 5. Linguistics research in China has shown a good development trend over the past 20 years, which keeps increasing year by year.

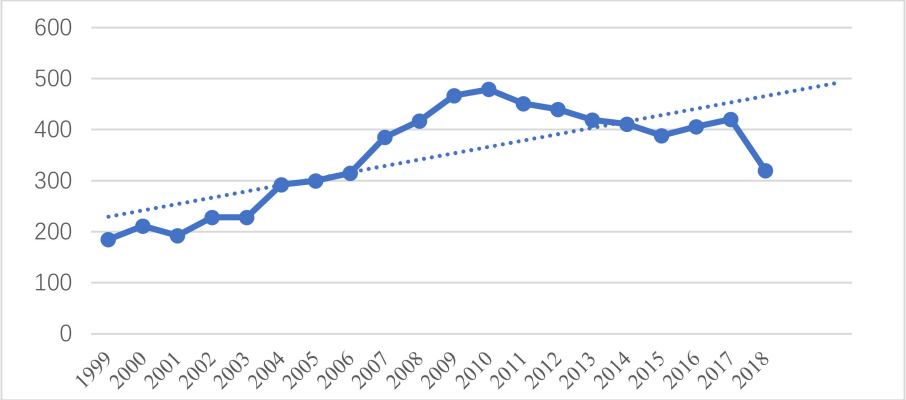


Figure 5. The trend of quantity change of linguistics periodicals in China.

This research extracts the titles, abstracts, and keywords of 6639 kinds of works and carries out word segmentation. Due to the particularity of linguistics, word segmentation software cannot accurately segment some professional terms, such as "applied linguistics", "comparative rhetoric", "computational linguistics", "systemic-functional linguistics", and so forth.

Hence, this paper takes the following steps. The first is the establishment of the User Dictionary. A total of 19,391 keywords were sorted out from the existing professional terminology database of linguistics and 6639 papers, and their parts of speech were marked as a noun (n.) to be introduced into the word segmentation system as a dictionary. Second, there is the extraction of nouns. The word segmentation results of each article were uniformly numbered as a unit, the nouns (marked as /n) were extracted by the Python programming language, and the meaningless nouns were automatically removed to obtain 19,834 nouns. The third step involves cleaning the 19,834 nouns, such as "role", "analysis", "research", and other meaningless nouns. Step 4 is noun translation. The data processing in step 3 needs to translate Chinese into English. Due to the differences in expressions between Chinese and English, we needed to consolidate many Chinese heterogeneous synonyms. For example, Chomsky and Noam Chomsky are corresponding words to the identical linguist. That aside, we marked the names of people, places, or countries with additional labels (i.e., the corresponding words directly added after "name", "city", and "country"), finally obtaining 9,183 English nouns. In step 5, we uniformly numbered the English words and replaced the nouns of the 6639 papers. Then, we retained the English terms when the word frequency in the corresponding articles was higher than 5 within the 1487 English words, according to the frequency of occurrence. At last, we set up the initial dictionary of linguistics research after sorting.

Due to the information noise, this paper combined the equivalence words and deleted ambiguous words and controlled the word frequency (>5) in the initial English dictionary (see Table 1), finally selecting 2111 representative words for this scientific research. At the same time, the chosen representative words were searched again to ensure that no literature information was lost so that these words could effectively represent the current situation of linguistics research in China. The data processing above was significant preparation for the subsequent feature maximization-based topic detection.

Table 1. Results of the dictionary processing procedure.

Procedure	Merge Equivalent Words	Delete Fuzzy Words	Control Word Frequency (>5)
Initial Data Size	9183	8845	8426
Disposed Words	——	419	6315
Merged Words	338	——	——
Final Data Size	8845	8426	2111

4. Research Design and Feature Maximization for Feature Selection

This study exploits a framework to explore the following subsections in Figure 6, which shows its complete data processing and the analysis. This study utilizes the articles' published times as a vital reference label to provide supplementary information for a more accurate understanding of the topic change.

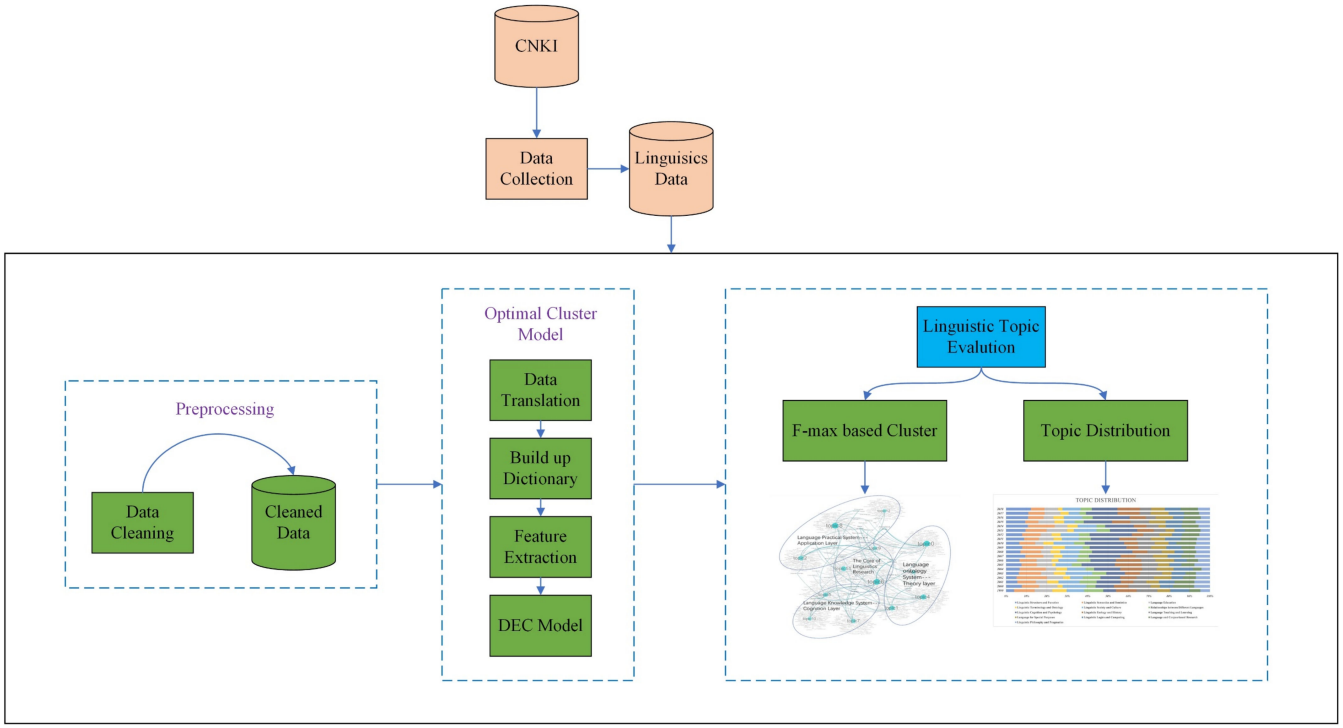


Figure 6. The research procedure of data analysis.

4.1. Clustering and Optimization Model Detection

Usual quality evaluators are sensitive to noise, so this paper further takes advantage of the *PC* and *EC* indexes proposed by Lamirel et al. [20] to make sure that the feature maximization algorithms and the contrast related to the activity and passivity of cluster features could effectively analyze high-dimensional data.

The *PC* index is mainly a macro-measure based on the maximization of the average weighted contrast of active features for the optimal partition. For a partition comprising k clusters, whose principle corresponds by analogy to that of intra-cluster inertia in the usual models, the *PC* index can be expressed as

$$PC_k = \underset{k}{\operatorname{argmax}} \left[\frac{1}{k} \sum_{i=1}^{k} \frac{1}{|s_i|} \sum_{f=S_i} G_i(f) \right] \tag{8}$$

Meanwhile, the *EC* index corresponds by analogy to that of the combination of the intra-cluster inertia and inter-cluster inertia in the usual models. The *EC* index is based on the maximization of the average weight compromising the contrast of active features

and the inverted contrast of passive features for the optimal partition. For a partition comprising k clusters, it can be expressed as

$$EC_k = \underset{k}{\operatorname{argmax}} \left[\frac{1}{k} \sum_{i=1}^{k} \left(\frac{|s_i| \sum_{f=S_i} G_i(f) + |\bar{s}_i| \sum_{h \in S_i} \frac{1}{G_i(h)}}{|s_i| + |\bar{s}_i|} \right) \right] \tag{9}$$

where n_i stands for the amount of data associated with the cluster i, $|s_i|$ is the number of active features in i, and $|\bar{s}_i|$ represents the number of passive features in the same cluster. Both indexes would help to obtain the optimal cluster number, which was also proved to be valid by Yue Chen et al. [24]. When comparing the *PC* and *EC* indexes, the *EC* index is especially suitable for the processing of high-dimensional text data. Notwithstanding the advantage of the *EC* index, it has not been used to determine and evaluate the clustering quality of linguistics research topics.

In our study, the *PC* values and *EC* values corresponding to 1–30 clusters were measured. (Because a single cluster is meaningless, this paper abandons the model with a cluster number of 1.) The dataset selected the clustering scheme according to the comparison of the *PC* and *EC* indexes and selected the model with 13 clusters as the optimal model (see Figure 7), which corresponded to the peak of the *EC* curve corresponding to the highest *EC* index value (i.e., the optimal contrast).

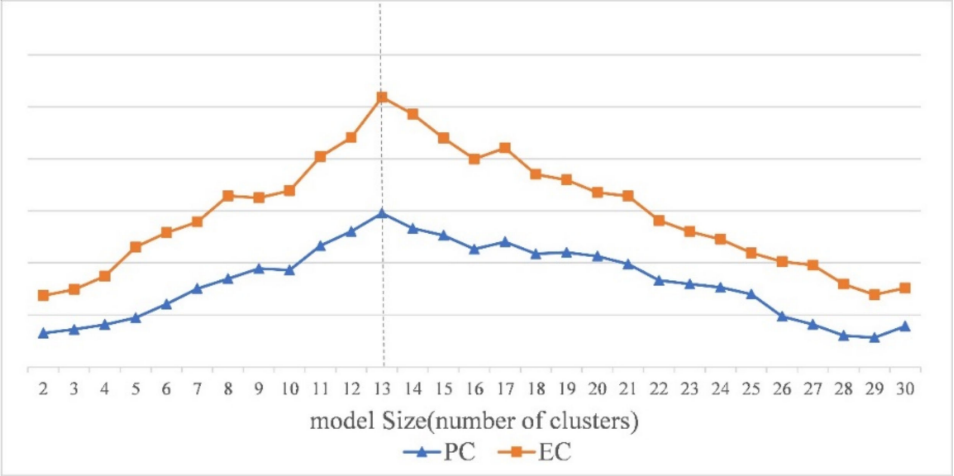


Figure 7. Trends of *PC* and *EC* indexes on linguistics research topics in the dataset of China.

Figure 7 draws the trends of the *PC* and *EC* indexes' evolutions in the case of linguistics research topics in China. It indicates what an appropriate *EC* index behavior is, while what describes the out-of-range index behavior was previously mentioned with the *PC* index in a parallel way. Moreover, the *EC* index was found to have more stable behavior under the noise sensitivity analysis circumstance.

Therefore, this method can scientifically optimize the number of linguistics research topics over time. After calculations and observations, the clustering labels were given according to the descriptive feature words of 13 clusters (see Table 2). In this step, each label tag was easily identified due to the feature maximization which extracted the active feature words. Additionally, the application of this combination based on the feature maximization in this research facilitated our unsupervised clustering. It provided a guarantee for the accuracy and stability of the clustering results in this paper.

Table 2. Active feature words and topic labels of the optimal clustering model.

Clustering	Label Name	Content (Active Feature Words)
Topic 0	Linguistic Structure and Function	System Function, Functional Structure, Systemic Functional Linguistics Theory, Interpersonal Function, Functional Discourse Analysis, Thematic Structure, Leonard Bloomfield, Textual Function, American Structuralism, Systemic Functional Grammar, Construction Grammar Theory, Functional Theory, Grammatical Unit, Functional Grammar Theory, Grammatical Pattern
Topic 1	Linguistic Semantics and Semiotics	Russian Semiotics, Symbolic Value System, Symbolic Arbitrariness, Linguistic Semiotics, Genre Study, Utterance Meaning, Signifier, Signified, Semantic Fuzziness, Semantic Processing, Textual Meaning, Sound and Meaning, Text Interpretation, Later Wittgenstein, Cultural Semiotics
Topic 2	Language Education	Teaching of Language and Literature, Grammar Teaching, Linguistic Teaching, Natural Language Understanding, Teacher Education, Language Environment, Discourse Teaching, English Reading Teaching, Bilingual Education, Chinese Language Studies, English Writing Teaching, Cooperative Principle, Educational Research, Curriculum Setting, Construction Theory, Chinese Education
Topic 3	Linguistic Terminology and Ontology	Ontology of Knowledge, Linguistic Terminology, Case-Auxiliary Word, Sense of Meaning, Theory of Meaning, Word Meaning, Social Turn, Meaning Potential, Terminology Translation, Taboo Words, Brand Naming, Interpreting Studies, Translated Names, Kinship Terminology, Dictionary Definition
Topic 4	Linguistic Society and Culture	Speech Community Theory, Discourse Style, Chinese Sociolinguistics, Language Variation, Identity Construction, Sociolinguists, Social Varieties, Language Contact, Politeness Principle, Socio-Cultural Factors, Cultural Difference, Cross-Cultural Communication, Register Analysis, Language and Society, Conversational Implicature
Topic 5	Relationships between Different Languages	Language Renaissance, Sino-Tibetan Languages, Tibetan, Burmese, Endangered Language, Language Comparison, Geographical Linguistics, Word Families in Chinese, Modern Chinese Dialect, Minority Language, Indo-European Languages, Language Family, Language World View, Ethnolinguistics, Language Evolution

Table 2. *Cont.*

Clustering	Label Name	Content (Active Feature Words)
Topic 6	Linguistic Cognition and Psychology	Concept Mapping, Conceptual Metaphor Theory, Mental Space, Metaphorical Meaning, Metaphorical Language, Multimodal Metaphor, Cognitive Linguists, Cognitive Metaphor, Cognitive Category, Conceptualization, Conceptual Integration Theory, Cognitive Schema, Cognitive Grammar Theory, Conceptual Representation, Cognitive Processing
Topic 7	Linguistic Ecology and History	Linguistic Ecology, Eco-Discourse Analysis, Language Ecosystem, Eco-linguistics, Historiography, Language Ecological Environment, Historical Narration, Immanence Theory, Language Diversity, Deep Structure, Surface Structure, Historical Research, Metalanguage, Philosophy of History, History of Rhetoric
Topic 8	Language Teaching and Learning	Linguistic Competence, Individual Difference, Learning Motivation, College English Teaching Model, Foreign Language Teaching, Teaching Method, Language Testing Theory, Second Language Acquisition Process, Teaching Strategies, Computer Assisted Instruction, College English Teaching Reform, Teaching Effectiveness, Foreign Language Learning, Applied Linguistics Theory, Autonomous Learning
Topic 9	Language for Special Purposes	Explanatory Turn, Legal Linguistic Psychology, Task-Based Language Learning, Economics of Language, Slogan, Business English, Tea Culture, Linguistic Nationality Studies, Tourism English, Language Taboo, Artistic Language, Culture Teaching, Manchu Script, Categorization Theory, the Use of Language, Target Language, Wittgenstein, Cultural Connotations
Topic 10	Linguistic Logics and Computing	Cognitive Logic, Complex Theory, Structural Law, Montague Grammar, Dependency Tree, Logic Language, Computational Simulation, Computational Linguistics, Vague Language, Decode, Language and Thinking, Mathematical Logic, Cognitive Neuroscience, Natural Language Processing, Machine Learning
Topic 11	Language and Corpus-Based Research	Chinese Corpora, Multimodal Corpus, Political Text, Sign Language Studies, Corpus-Based Translation Studies, Corpus Stylistics, Complex Network, Corpus Linguistics, Word Frequency, Data Mining, Discourse Coherence, Advertising Language, Pragmatic Features, Corpus Approach, Chinese Information Processing

Table 2. *Cont.*

Clustering	Label Name	Content (Active Feature Words)
Topic 12	Linguistic Philosophy and Pragmatics	Discourse View, Dialogue Theory, Discourse Theory, Communication Strategy, Discourse Markers, Pragmatic Turn, Philosophical Thought of Language, Linguistic Philosophy, Conversational Implicature, Legal Discourse, Philosophical Thinking, Discourse System, Pragmatics Research, Cognitive Pragmatics, Pragmatic Inference

Note. The clustering contrast value of Topic 6 (Linguistic Cognition and Psychology) was the largest, and the clustering feature words and their F-values were as follows: 13 for Concept Mapping, 13 for Conceptual Composition Theory, 13 for Conceptual Metaphor Theory, 13 for Mental Space, 13 for Metaphorical View, 13 for Ontology Concept, 13 for Synecdoche, 12.106781 for Metaphorical Meaning, 12.042655 for Metaphorical Language, 11.889931 for Multimodal Metaphor, 11.83.7066 for Cognitive Linguists, 11.526544 for Frame Theory, 11.398952 for Cognitive Metaphor, 11.344811 for Cognitive Category, 11.273071 for Conceptualization, 11.200615 for Conceptual Integration Theory, 10.953567 for Conceptual Metaphor, 10.889600 for Cognitive Tools, 10.832169 for Cognitive Schema, 10.493149 for Prototype Category Theory, 10.475425 for Metaphor, 10.449791 for Embodied Philosophy, 10.441923 for Metaphor Research, 10.418085 for Metaphor Theory, 10.271934 for Metaphtonymy, 9.4058784 for Cognitive Grammar Theory, 9.1963083 for Cognitive Experience, 8.877107 for Conceptual Structure, 8.796274 for Metaphorical Thinking, 7.6359203 for Cognitive Semantics, 7.0802880 for Conceptual Representation, 6.55355658 for Cognitive Processing.

4.2. Contrast Graph and Its Representation

The contrast graph is a bipartite graph based on the relationship between feature set S and label set L [36]. The bipartite graph connects two independent sets U and V, and the two sets do not intersect with each other, with one edge connecting the nodes. Theoretically, the label set L could express various information about the associated features, and the feature set S, a subset of feature set F, was obtained through the feature selection process. When using the feature maximization algorithms, the weight $c_{(u,v)}$ of edge (u, v), $u \in S$, $v \in L$ represents the contrast of feature u of label v. The labels in this study were abstracted from the relevant data of clustering.

This paper hereafter utilizes Cytoscape software to draw the bipartite graph. This graph has the following three features: (1) the number of connections is appropriately reduced in the process of correlation feature selection to alleviate the cognitive overload caused by graph representation, (2) when feature words are connected with multiple labels, they can show the relationship between labels, and (3) combining this method with the weighted orientation model and visualization can highlight the core of the most influential labels in the label L set, and the feature words closely related to the label will gather in the adjacent position of the label. Subsequently, the obtained clustering topics would use this technique for clear illustration.

5. Results and Discussion

We designed a research procedure to extract the linguistics research topics in China over time based on the feature maximization and a high-dimensional data clustering algorithm.

5.1. Topic Clustering for the Linguistics Domain

According to the feature maximization and the optimal clustering model, this study obtained 13 clustering topics and retained 1487 feature words with F-values higher than 3. The contrast graph represents the topic structure of linguistics research by way of visualization (see Figure 8.). This article exploits an optimal clustering model with the combination of feature maximization and the contrast ratio as well as the DEC clustering algorithm to effectively optimize the number of topic clusters. Due to the optimal partition mentioned above, it was expected to maximize the contrast described by Equation (6). It was found that the higher the feature contrast was, the greater the compactness within the class, and the higher the discrimination between classes was. This paper hereafter uses the

contrast ratio to perform clustering visualization, which not only solves the problem of cognitive overload of the interactive representation of large datasets but also extracts and displays the connections between topics through high-contrast shared features.

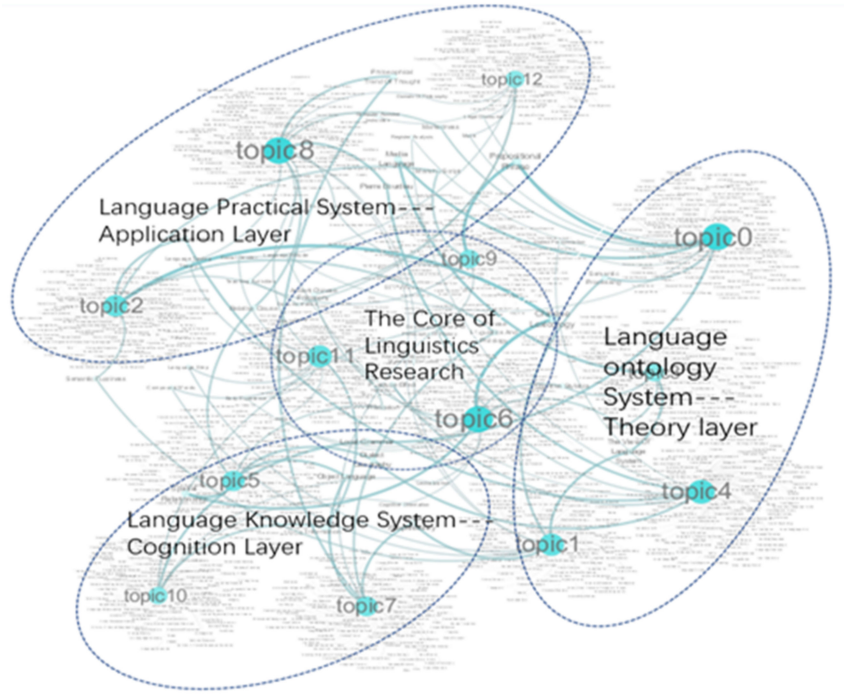


Figure 8. A structural map of linguistic research topics in China. Note: see Appendix A below for a partial enlargement of the image, which highlights topic 6: linguistic cognition and psychology.

As far as any discipline is concerned, it is inseparable from language as a carrier. Over the past 20 years, facing the development of linguistics research itself, the study of language cognition, how human language reflects thought, and psychological behavior is regarded as a core topic of linguistics. Additionally, the study of language for special purposes is another important core topic that mainly investigates various linguistic genres (e.g., news reports, legal contracts, experiment reports, and dissertations) to meet the specific needs of language research in all walks of life. Thus, in the context of big data, the traditional introspective method no longer satisfies the requirements of language research at present. Therefore, corpus resources and linguistic analyses as the future methodological direction make available a broader range of languages in language science. Such resources and analyses can play a transformative enabling role in testing and developing theories of language structure based on the principles of efficiency, learnability, and formal parsimony [37]. Consequently, the topic of language and corpus-based study evolves into a hotspot, which principally focuses on systematic language research based on corpora. It depends on natural language processing and text detection to scientifically extract the law of language.

According to Figure 7, we obtained 13 topics distributed as follows. Figure 8 suggests that Topic 6 ("linguistic cognition and psychology"), Topic 9 ("language for special purposes"), and Topic 11 ("language and corpus-based study") lie in the core location among the 13 topics in the linguistics domain. Around the three core topics, the linguistics domain is supported by three major areas, namely the practical language system, language ontology system, and language knowledge system, which are equal to the application layer, theory layer, and cognition layer, respectively, constituting a complete logical research system for linguistics.

(1) On the linguistic theoretical layer, the domain of the "language ontology system" contains four related topics: Topic 0 (language structure and function), Topic 1 (linguistic semantics and semiotics), Topic 3 (linguistic terminology and ontology), and Topic 4

(linguistic society and culture). The four clustering topics comprehensively reflect the domain of the language ontology system from the linguistic theoretical layer. Language is one of the most important communication tools for human beings. Its function is a way to express and communicate their ideas, feelings, and desires. The form of language itself is also a symbol system; otherwise, language will not spread without social history and culture. The meaning of language exists in the process of people's understanding of its application and practice. A nation will integrate its cognition of the objective world into its language habits. The combination of linguistic meaning and symbols is essentially a response to this fact and the view of the world, and it is also a kind of network structure. It is the representation form of different levels or types of a language ontology system. It acts as a structural output between a language system and context (i.e., the correlation of language structural elements). The structural semantics of the language are based on the recognition of the qualitative social characteristics of language signs. As for language philosophy from ontology to epistemology, the ontological study of the universal symbol is performed through the special method of hermeneutics. From the dimension of ontology and hermeneutics, the structure and meaning of symbols are an integral relationship in meaning, form, and content. The symbol acts as a carrier both inside and outside the dimension of human cognition. Language is a specific part of speech activities and a symbolic system to express ideas through terms and ontologies. Ontologies can represent knowledge in specific domains and enable semantic interoperability by being connected to other external data sources [38]. Additionally, language, reflecting all human cultural phenomena, is a symbol of social developments and changes. Regarding language from tools in the ontology, language itself has become the object of theoretical investigation and regarded as the starting point of theory. The study of language ontology cannot be separated from the sociality of language signs. The study of a language's social attributes is a conscious and systematic study of language, a unique social system for maintaining society with the function of language organization.

(2) On the linguistic cognitive level, the field of "language knowledge system" highlights three related topics: Topic 5 (the relationship between different languages), Topic 7 (linguistic ecology and history), and Topic 10 (linguistic logic and computing). Language is man's cognitive boundaries, and language itself is life. The relationships between languages focus on investigating how languages of various language families and affinities interact with and transform each other. Language is the unique gift of human beings which endows language with the nature of all organic life. The law of language development is similar to the evolutionary process of living things. The diversity of languages, endangered languages, and human rights of languages are the hot spots of language and ecology research. Language as a carrier can witness changes in social history, and language itself changes in the dimension of time and space correspondingly. Aside from that, the language has not only biological properties but also mathematical properties. It reflects the content of thinking through the method of language processing in meaning and sound. Topic 10 is interested in how people organize their thoughts into language to avoid errors caused by semantic ambiguity, semantic contradictions, or structural confusion. In essence, language elements can code and generate unlimited meaning with limited units and limited rules in the mathematically logical way. Machine learning and artificial intelligence, as emerging fields, are closely related to language logic and computation now. Therefore, the study of natural language structure and behavior has become a hotspot of the language knowledge system.

(3) On the linguistic application layer, three related topics support the field of "practical linguistic system", including Topic 2 (language education), Topic 8 (language teaching and learning), and Topic 12 (linguistic philosophy and pragmatics). Language as a communicative medium promotes human beings to communicate with each other, conduct their activities in society, and understand the world in the application layer. Topic 2 and Topic 8 are the basic representations of the language practice system. Language teaching and learning highlight how people can acquire language-related knowledge and strategies

of using language. These examples of research pay special attention to the linguistic and ecological diversity of language learning. That aside, both topics focus on how individuals compare the similarities and differences between known and unknown languages in the information process of language learning and cognition. Language teaching and learning strategies are lasting hotspots. Regarding Topic 2 (language education), this topic closely interacts with teacher education, linguistic educational research, and the language environment, which are mainly associated with language knowledge and application. In particular, pragmatics comes from the philosophy of language and has brought significant attention from different research communities because it mainly investigates the use of specific language in different contexts, the output, and the understanding of utterances. Pragmatics turns to a new platform of philosophical dialogue with the achievements of philosophy, which is partly due to ordinary language philosophy being a kind of pragmatic philosophy. The meaning of a language lies in its use and is no longer predetermined but revealed in the net of the acts of use. Therefore, Topic 12 is an indispensable research topic in the practice system of the language, which focuses on the relationship between language and the world.

5.2. The Evolution of Linguistics Research Topics

According to the 13 clustering topics of linguistics research in the past 20 years, both Figures 9 and 10 clearly show their historical paths of change. Since the end of the 20th century, linguistic research in China has been thriving. Linguistic research started from the study of language ontology, with many topics centering on the internal features of the language. In 1999, the academic circle mainly discussed the internal features of different languages and investigated the relationship between languages from the perspective of multi-language comparison and contrasting. Since 2001, the academic community has developed a strong interest in the language ontology system and knowledge system. In 2001, it focused on the study of language terms and ontology. In 2002, the major research topic of semantics and semiotics covered all levels of language research, which provided an important theoretical basis for linguistic research. From 2003, linguistic theories on interacting with other disciplines began the interdisciplinary research. In 2003, the focus was put on linguistic interaction generated by the sociolinguistics and sociology theory, a hot topic in this period mainly researching the relationship between language, society, and culture. In 2004, linguistics, logic, and computer technology combined to generate computational linguistics. Topic evolution is the incremental change of either a feature space (i.e., the composition of the involved terms) or data distribution (i.e., the frequency of associated terms) in a topic, and such a change results in the appearance of new topics [39]. The evolution of linguistics research topics is the trend of topics' vicissitude in a broader historical time. As is known to us, language study is a bridge between humanities, social sciences, and natural sciences. Linguistics has become a leading subject in the fields of philosophy, computer applications, knowledge engineering, and so on.

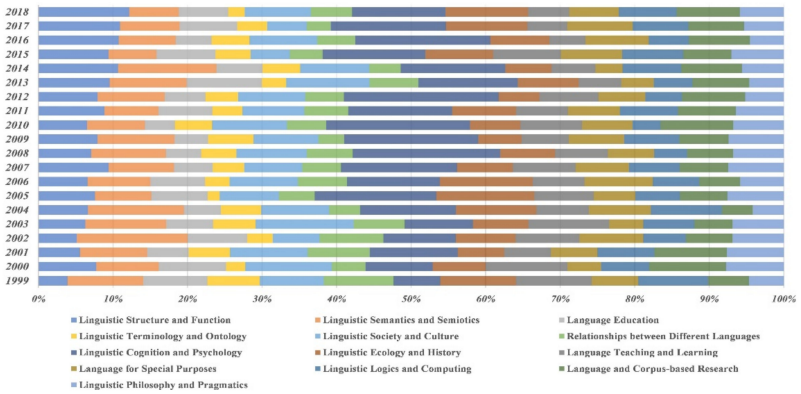


Figure 9. The topic distribution according to the integrated approach.

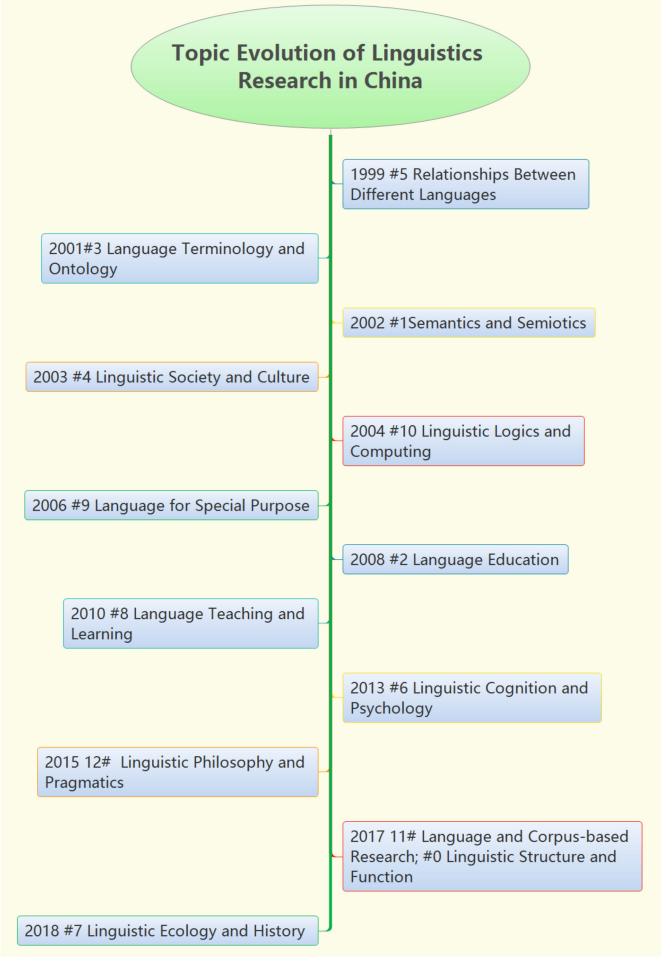


Figure 10. The topic evolution of linguistics research in China.

Subsequently, linguistics research entered a phase of rapid development, and the number of published papers increased year by year, together with the topics highlighting the application layer of language from 2006 to 2010. In addition, the number of published articles was the largest in 2010 among these 5 years. The three research topics, including special purpose language in 2006, language education in 2008, and language teaching and learning in 2010, confirm the high focus of linguistics in language application, indicating that the academic circle attaches great importance to the linguistic instrumentality.

After 2013, the change in linguistics topics turned to a new phase. Five major topics commonly emphasize that language as a system needs investigating from the perspective of system theory, namely Topic 6 (linguistic cognition and psychology) in 2013, Topic 12 (linguistic philosophy and pragmatics) in 2015, Topic 11 (language and corpus-based study), Topic 0 (linguistic structure and function) in 2017, and Topic 7 (linguistic ecology and history) in 2018. These topics indicate that language research gradually developed from experience in the introspection method to the objectively empirical study [40]. In particular, the main foci of this period were closely related to various linguistic theories. The quantitative study aims at exploring new language rules from theory to practice and then from practice to theory again to obtain the sublimation of language cognition. Among them, the influences of corpus linguistics, cognitive linguistics, and ecolinguistics are expanding rapidly, which indicates that they will maintain steady development momentum in the future. Linguistics is an open discipline which not only sheds light on the study of language ontology but also combines with other disciplines' theories and technology to form a subject tightly integrated with characteristics, interdisciplinary factors, and practice.

The insights presented in this article are of great importance to not only scientific researchers but also editors for academic journals. The findings could do us a favor in

understanding linguistics research's evolutionary history and its research status, as well as distinguish between prevalent and declining topics in linguistics research. Additionally, the quantitative method based on the F-maximization index, the contrast ratio measurement algorithms, and the DEC clustering algorithm based on the KTRM could bring a new perspective of the research methodology for humanities and social science research. Hence, our results can be used to guide future research activities. Moreover, researchers in linguistics and other interdisciplinary fields could adjust the scope of their research topics to prioritize research hotspots or pay more attention to topics that are, in fact, significant. Research on these and other topics should help us gain new and more in-depth understandings of the issues involved in language learning, use, and communication in general [41].

5.3. Predicting the Trend of Hotspots in Linguistics Research

This study generates a word cloud map by Python programming to represent the hot topics in the linguistics domain. According to the word frequency of each feature word in the collected dictionary, the size of the font represents the number of keywords that appear in the collection. In principle, the font size of a word in the word cloud is determined by its appearance frequency. We applied a China map-like mask and "impact" font, and then we obtained a graphical picture. The illustration is shown in Figure 11.

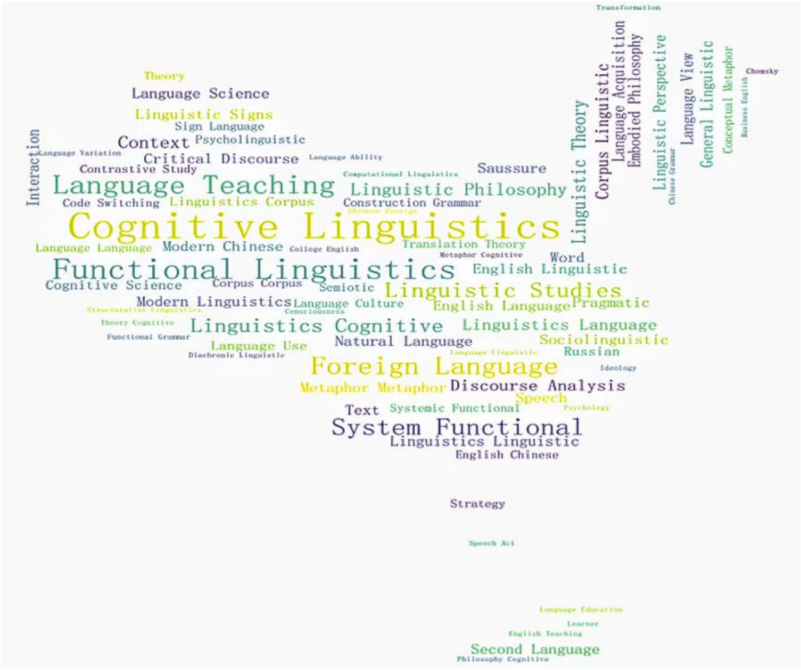


Figure 11. Word cloud map of hotspots in the linguistics domain.

Figure 11 indicates that linguistic cognition, linguistic function, language teaching, metaphor, corpus linguistics, discourse analysis, and linguistic philosophy are the hotspots in future linguistics studies. From the perspective of research hotspots, the study of language function and a symbol system is an everlasting hotspot in linguistics. Aside from that, with the rapid development of the information age, language research tends to adopt scientific data mining methods. Therefore, corpus linguistics and computational linguistics have thrived in recent years. The symbolic function of language has been an intriguing hotspot in cognitive science research, too. The achievements of linguistics research in China in the field of cognitive science continue to increase, and its influence also continues to grow. That aside, critical discourse analysis and natural language processing are hotspots in linguistics, followed by linguistic philosophy. Language ecology and endangered languages are the keywords with high word frequencies, which are also the research hotspots in the past five years and even extend to future studies. After addressing

the concerns of the questions above, we pushed this study forward by predicting the future trends of linguistics research in China. Word clouds, a kind of weighted list to visualize language or text data, have gained increasing attention and more application opportunities as a big data approach [20].

6. Conclusions

This paper applies the combinatory approach with the F-index, the contrast ratio, and the DEC clustering algorithm based on the KTRM method to detect the new yet increasing research stream on the change in linguistics research topics in China. Our study contributes to expanding the application of topic detection based on the feature maximization and *EC* index to address the change in topics in linguistics research and the hotspots in various research phases. Through this method, our findings reveal that linguistics research focuses on three core topics, and other topics extend to the following three main linguistic research layers around the core topics: the linguistic ontology system, the linguistic knowledge system, and the linguistic application system. Each layer covers vital topics that could review the evolution of linguistics research topics. Working in this study has infused linguistics study with new precision and methodological rigor. Our study offers insights into the functional role of the combinatory approach.

In addition, this study contributes to the efforts to broaden the research horizon for researchers not only in the field of linguistics but also in the interdisciplinary domains. From the results of 13 clustering topics, we found that the research scope of linguistics has extended to studies in various main topics and represents a trend of the increasingly interdisciplinary feature, especially in cognitive science and computational science. The findings would be beneficial for researchers, journals, and publishers in finding intriguing hotspots and topics for future research. This study demonstrates that the combinatory method is useful and time-saving when conducting high-dimensional classification problems without clustering parameter estimation. Therefore, the advantages of this combinatory method can provide a reference for analyzing the topic evolution of other domains.

However, there are still some limitations to this research. We only detected the change in linguistics research topics in China and neglected those of other countries. We only adopted *EC* values to confirm the amount of clustering and neglected comparing those of other estimate indexes. Aside from that, we only applied this combinatory method to explore topic detection in the linguistics domain rather than conducting further exploratory studies in different fields. Hence, further experiments are required that use both an extended set of clustering methods and a larger panel of high-dimensional datasets of linguistics research in foreign countries to confirm this *EC* value's behavior and the related quality estimators. Additionally, we plan to apply this approach to examining the change in other domains of research over time to provide innovative research perspectives and a methodological reference for future diverse domain knowledge research.

Author Contributions: Conceptualization, J.F. (Junchao Feng) and Y.T.; methodology, Y.L.; software, Y.L.; validation, J.F. (Jundong Feng); formal analysis, J.F. (Junchao Feng); investigation, Y.T.; resources, Y.L.; data curation, Y.T.; writing—original draft preparation, J.F. (Junchao Feng); writing—review and editing, J.F. (Junchao Feng); visualization, J.F. (Junchao Feng); supervision, J.M.; project administration, J.F. (Jundong Feng). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [Grant Number 11705089], the Fundamental Research Funds for the Central Universities [Grant Number NS2021038], the Key Research Project for Economic and Social Development in Heilongjiang Province China [Grant Number [WY 2021054-C], and the 2021 Heilongjiang Provincial Philosophy and Social Science Research General Project [Grant Number. 21YYB163].

Institutional Review Board Statement: Not applicable.

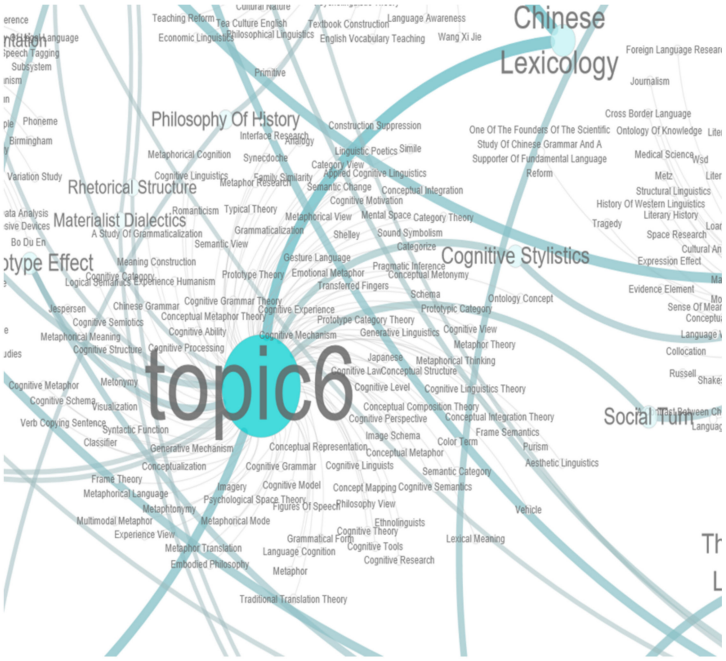
Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank Li Xiaofeng for his research suggestion, and the authors quite appreciate the editor of the journal and the anonymous reviewers. Their constructive, insightful comments and suggestions have helped significantly enhance the quality of this article.

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

Appendix A



References

1. Li, F.; Li, M.; Guan, P.; Ma, S.; Cui, L. Mapping Publication Trends and Identifying Hot Spots of Research on Internet Health Information Seekinsg Behavior: A Quantitative and Co-Word Biclustering Analysis. *J. Med. Internet Res.* **2015**, *17*, e81. [CrossRef] [PubMed]
2. Lamirel, J.-C. A new approach for automatizing the analysis of research topics dynamics: Application to optoelectronics research. *Scientometrics* **2012**, *93*, 151–166. [CrossRef]
3. Neshati, M.; Fallahnejad, Z.; Beigy, H. On dynamicity of expert finding in community question answering. *Inf. Process. Manag.* **2017**, *53*, 1026–1042. [CrossRef]
4. Hu, K.; Luo, Q.; Qi, K.; Yang, S.; Mao, J.; Fu, X.; Zheng, J.; Wu, H.; Guo, Y.; Zhu, Q. Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Inf. Process. Manag.* **2019**, *56*, 1185–1203. [CrossRef]
5. Zhang, X. A bibliometric analysis of second language acquisition between 1997 and 2018. *Stud. Second Lang. Acquis.* **2019**, *42*, 199–222. [CrossRef]
6. Garcia, K.; Berton, L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl. Soft Comput.* **2021**, *101*, 107057. [CrossRef]
7. Chen, W.; Chen, W. The identification and evolution of research frontiers from comparison of science and technology. *J. Intell.* **2022**, *41*, 67–73, 163.
8. Chen, X.; Wang, S.; Tang, Y.; Hao, T. A bibliometric analysis of event detection in social media. *Online Inf. Rev.* **2019**, *43*, 29–52. [CrossRef]
9. He, Q.; Chang, K.; Lim, E.P. Analyzing feature trajectories for event detection. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 207–214.
10. Ding, Y. Community detection: Topological vs. topical. *J. Inf.* **2011**, *5*, 498–514. [CrossRef]
11. Li, S.; Li, M. A new paradigm of interdisciplinary research on linguistics: A review of PANS research from 2000 to 2016. *Lang. Teach. Linguist. Stud.* **2019**, *1*, 102–112.
12. Duan, L.; Ma, S.; Aggarwal, C.; Sathe, S. Improving spectral clustering with deep embedding, cluster estimation and metric learning. *Knowl. Inf. Syst.* **2021**, *63*, 675–694. [CrossRef]
13. Kim, J.; Yoon, J.; Park, E.; Choi, S. Patent document clustering with deep embeddings. *Scientometrics* **2020**, *123*, 563–577. [CrossRef]

14. Kassab, R.; Lamirel., J.C. Feature Based Cluster Validation for High Dimensional Data. In Proceedings of the International Conference on Artificial Intelligence and Application, Innsbruck, Austria, 2 June 2008; pp. 97–103.
15. Dayeen, F.R.; Sharma, A.S.; Derrible, S. A text mining analysis of the climate change literature in industrial ecology. *J. Ind. Ecol.* **2020**, *24*, 276–284. [CrossRef]
16. Shen, S.; Li, Q.Y.; Ye, Y.; Sun, H.; Ye, W.H. Topic Mining and Evolution Analysis of Medical Sci-Tech Reports with TWE Model. *Data Anal. Knowl. Discov.* **2021**, *5*, 35–44.
17. Mustak, M.; Salminen, J.; Plé, L.; Wirtz, J. Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *J. Bus. Res.* **2021**, *124*, 389–404. [CrossRef]
18. Coppens, F.; Wuyts, N.; Inzé, D.; Dhondt, S. Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Curr. Opin. Syst. Biol.* **2017**, *4*, 58–63. [CrossRef]
19. Chen, M.; Flowerdew, J. Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *Engl. Specif. Purp.* **2018**, *50*, 97–112. [CrossRef]
20. Liu, H.; Lin, Y. Methodology and Trends of Linguistic Research in the Era of Big Data. *J. Xinjiang Norm. Univ. (Philos. Soc. Sci.)* **2018**, *1*, 72–83.
21. Liu, Y. Information Visualization Analysis on the Research Hot Spots and Frontiers of International Corpus Linguistics. *Knowl. Manag. Forum* **2018**, *3*, 208–224.
22. Li, Z.; Xu, J. The evolution of research article titles: The case of Journal of Pragmatics 1978–2018. *Scientometrics* **2019**, *121*, 1619–1634. [CrossRef]
23. Lamirel, J.C.; Dugué, N.; Cuxac, P. New efficient clustering quality indexes. In Proceedings of the International Joint Conference on Neural Networks IEEE (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 3649–3657.
24. Chen, Y.; Lamirel, J.-C.; Liu, Z. An overview on 40 years science of science research topic evolution in China: A novel approach based on clustering and feature maximization. *Sci. Sci. Manag. Sci. Technol.* **2018**, *39*, 28–45.
25. Xie, J.; Girshick, R.B.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 24 May 2016; Volume 48.
26. Pan, Y.; Wang, M.; Wang, J. Clustering of agricultural trade friction news text based on improved text representation and its application prospect. *Agric. Outlook* **2020**, *16*, 80–88.
27. Chen, B.; Tsutsui, S.; Ding, Y.; Ma, F. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *J. Inf.* **2017**, *11*, 1175–1189. [CrossRef]
28. Hui, L.; Jixia, H.; Zhiying, T. Subject topic mining and evolution analysis for multi-source data. *Data Anal. Knowl. Discov.* **2022**, *31*, 1–16.
29. Mane, K.K.; Börner, K. Mapping topics and topic bursts in PNAS. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5287–5290. [CrossRef] [PubMed]
30. Blei, D.M.; Lafferty, J.D. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 113–120.
31. Lounsbury, J.W.; Roisum, K.G.; Pokorny, L.; Sills, A.; Meissen, G.J. An analysis of topic areas and topic trends in the Community Mental Health Journal from 1965 through 1977. *Community Ment. Health J.* **1979**, *15*, 267–276. [CrossRef]
32. Lamirel, J.-C.; Francois, C.; Al Shehabi, S.; Hoffmann, M. New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. *Scientometrics* **2004**, *60*, 445–562. [CrossRef]
33. Lamirel, J.C.; Mall, R.; Cuxac, P.; Safi, G. Variations to incremental growing neutral gas algorithm based on label maximization. In Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN), San Jose, CA, USA, 3 October 2011; pp. 956–965.
34. Lamirel, J.-C.; Cuxac, P.; Chivukula, A.S.; Hajlaoui, K. Optimizing text classification through efficient feature selection based on quality metric. *J. Intell. Inf. Syst.* **2015**, *45*, 379–396. [CrossRef]
35. Dempster, A.P. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
36. Cuxac, P.; Lamirel, J.C. Analysis of evolutions and interactions between science fields: The cooperation between feature selection and graph representation. In Proceedings of the 14th COLLNET Meeting, Tartu, Estonia, 14 August 2013; pp. 780–788.
37. Futrell, R.; Mahowald, K.; Gibson, E. Large-scale evidence of dependency length minimization in 37 languages. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 10336–10341. [CrossRef] [PubMed]
38. Zhang, J.; El-Diraby, T.E. Social semantic approach to support communication in AEC. *J. Comput. Civ. Eng.* **2012**, *26*, 90–104. [CrossRef]
39. Zhang, Y.; Chen, H.; Lu, J.; Zhang, G. Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowl.-Based Syst.* **2017**, *133*, 255–268. [CrossRef]
40. Lei, L.; Liao, S. Publications in Linguistics Journals from Mainland China, Hong Kong, Taiwan, and Macau (2003–2012): A Bibliometric Analysis. *J. Quant. Linguist.* **2017**, *24*, 54–64. [CrossRef]
41. Jin, Y. Development of Word Cloud Generator Software Based on Python. *Procedia Eng.* **2017**, *174*, 788–792. [CrossRef]