Article

Not peer-reviewed version

A Method for Identifying Proper Noun Derivative Semantic Relationships for Geographic Entities

Liu Hanyou and Wang Jizhou *

Posted Date: 15 April 2024

doi: 10.20944/preprints202404.0941.v1

Keywords: derivation relationship of proper nouns between toponyms; proper nouns derive toponym; extraction of derivatives of proper nouns; the morphological characteristics of toponym; spatial and temporal distribution characteristics of toponym

 doi:10.20944/preprints202404.0941.v1

Article

A Method for Identifying Proper Noun Derivative Semantic Relationships for Geographic Entities

Hanyou Liu [1] and Wang Jizhou [2,*]

[1] Liaoning Technical University
[2] Chinese Academy of Surveying and mapping
* Correspondence: wangjz@casm.ac.cn

Abstract: There are rich proper noun derivative semantic relationships in geographic entity vector data. The identification of traditional proper noun derivative semantic relationships usually requires toponym researchers to examine the source of derived toponyms by consulting a large number of toponym documents. This method cannot adapt to the mining of proper noun derivative semantic relationships for large-scale geographical entity. In this regard, this paper proposes a method of identifying the proper noun derivative semantic relationships for geographical entity by combining prompt learning and XGBoost model. This method transforms the extraction problem of proper noun derivative semantic relationships into a binary classification problem by analyzing the characteristics of the derivation relationship of toponyms in the form of toponyms and the spatial and temporal distribution of geography. Firstly, prompt learning and gaussian mixture model are used to extract and cluster the derivation words of toponyms, so as to realize the recognition of derived toponyms collections. Then, enumerate the possible proper noun derivations in the toponym set, and then the XGBoost model is used to identify the semantic relations derived from the proper noun of geographical entities. In the experiment, the recognition method of proper noun derivative semantic relationships based on prompt learning and XGBoost proposed in this paper has achieved good relationship recognition effect in the experiment. This method has certain practical value in the fields of toponym management and toponym translation.

Keywords: derivation relationship of proper nouns between toponyms; proper nouns derive toponym;extraction of derivatives of proper nouns; the morphological characteristics of toponym; spatial and temporal distribution characteristics of toponym

---

1. Introduction

The proper noun derivative relationship between the names of geographic entities and their derived toponym is a semantic relationship, such as the relationship between the geographic entity "New York City" and the geographic entity "New York University". It is a common method to use the toponym derivation of one native toponym to naming another new toponym, in the naming process of newborn geographic entities, people usually choose the original geographic entity name of the onomastic word or its derivatives, as the newborn geographic entity name of the onomastic part of the new geographic entity name, in order to express the newborn geographic entity and the original geographic entity of the geographic association relationship(such as the relationship of neighboring, subordinate relationship). At this time, the geographical name of the newborn geographical entity is called a derived toponym, and the geographical name of the original geographical entity is called a native toponym, and there is a toponym derived relationship between them. There are two main forms of onomastic derivation: 1) direct derivation, through the selection of the whole or part of the original geographical name of the onomastic part of the original geographical name directly as a derivative part of the original geographical name. For example, the toponym "Ontario Center Cemetery graveyard" and the original toponym "Ontario County", of which the derived part of the toponym of the onomastic derivation is "Ontario".2) Indirect derivation, in which the derivatives of the whole or part of the words of the native toponym proper are taken as derivatives, such as the derivation of the toponyms "Washingtonville Post Office" and "Washington County", where the derived part of the toponym is "Washingtonville", "Washingtonville" is a derivative of "Washington".

 doi:10.20944/preprints202404.0941.v1

The extraction of derivational relations between proper noun of geographic entities mainly involves two research fields: named entity recognition and derived toponyms.

In the field of named entity recognition, with the continuous development of large language models, the "pre-trained + prompted + predicted" paradigm based on prompt learning has gradually replaced the "pre-trained + fine-tuned" paradigm as the mainstream model of current natural language processing tasks. Unlike the "pre-trained + fine-tuned" paradigm in which pre-trained language models are designed according to downstream tasks, prompt learning trains a language model $P(y|x; promot; \sigma)$ through a datasets, where x is the input text, $prompt$ is the additional added context information, and y is the output text label. This method adds prompt information to the original input text, and guides the model to produce a specific output. In terms of information extraction, the difference between language model training and downstream tasks is shortened. Naming has done a lot of research on named entity extraction based on prompt.Siyu Yuan et al. (2023) using the topic information of entities to be extracted in the existing knowledge graph as prompt, combined with structural causal model, effectively improves the concept extraction effect of entities[1].He et al. (2023) proposed PWII-BERT (Prompt-based Word-level Information Injection BERT) method, which eliminates the concept bias caused by false co-occurrence relations by using word-level category information as cue words and combining lexicon features[2].Rozhkov et al. (2023) investigated the effects of different prompts on a few-sample nested named entity recognition task[3].Dhananjay et al. (2023) show that PromptNER in ConLL and FewNERD datasets outperforms existing methods in both small sample named entity recognition and cross-domain named entity recognition[4].Chen et al. (2023) aiming at the problem of named entity recognition with few samples in the medical field, a knowledge-guided instance generation method based on the domain knowledge graph is proposed. The named entity extraction problem is transformed into a question-and-answer task, and the question is used as a prompt. Then the comparative learning method is used to improve the generalization ability of the model[5].

In the field of derived geographical names,Hobo Cheng (2016) briefly analyzes and introduces the concept, characteristics and types of derived toponyms[6],Tan Ruwei (2018) summarizes common derived toponyms in china[7],Cosimo, Palagiano (2016) analyze and summarize the various sources of toponyms in the united states[8]. Liu Hanyou et al. (2022) analyzes the characteristics of english fully derived toponyms and common name derived toponyms, and then sets the corresponding conditional threshold by constructing multiple constraints to identify fully derived toponyms and common name derived toponyms[9].On this basis, Liu Hanyou (2022) constructs a knowledge graph of english derived toponyms to assist in the recognition of derived toponyms[10].

At present, in the field of information extraction, with the advent of large language models, the method based on prompt learning has gradually become the mainstream method of information extraction; in the field of derived toponyms, many scholars at home and abroad have done a lot of research work on the characteristics, types and recognition methods of derived toponyms, but the recognition of toponym derivation is still in a blank stage. The recognition of the toponym derivation relationship of geographical entities is of great significance in the fields of toponym translation and toponym research. However, at present, there are relatively few studies on the study of derived toponyms, mainly focusing on the study of complete derived toponyms and generic derived toponyms. Therefore, by analyzing the characteristics of the toponym derivation relationship, this paper adopts the method of supervised machine learning to extract the toponym derivation relationship between geographical entities from the vector data of geographical entities and the structured toponym attribute data.

2. Methods

In this paper, the method of pipeline is used to extract the toponym derivation relationship, it can be roughly divided into two steps: the recognition of place names derived toponym from proper nouns and the determination of the relationship derived from proper nouns.

 doi:10.20944/preprints202404.0941.v1

2.1. Recognition of Toponyms Derived from Proper Noun

From a linguistic point of view, on the word morphology of the proper noun part, the derived toponym and the original toponym have the same or similar proper noun. From the perspective of the toponym derivation relationship, there is a one-to-many relationship between the original toponym and the derived toponym, that is, the proper noun derivatives of multiple toponyms derived from the same original toponym come from the toponym of the same original toponym. Therefore, the derived toponyms derived from the same original toponym have similar or the same proper noun derivatives. In this paper, the strategy of "identification first, then clustering" is adopted. First, the proper noun derivatives are extracted from the sequence of toponym words to identify the derived toponyms; then, the clustering of the proper noun derivatives is used to realize the recognition of the corresponding set of derived toponyms. The recognition of toponyms derived from proper noun is mainly divided into two parts: the extraction of proper noun derivatives and the clustering of toponyms derived from proper noun.The overall flow is shown in Figure 1.

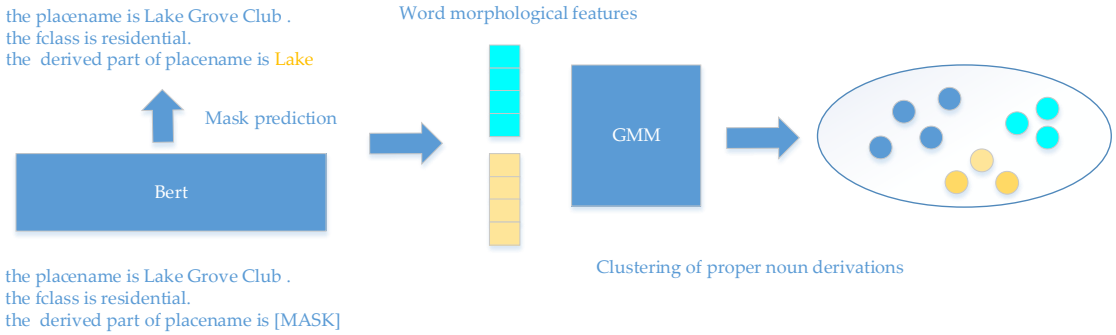

Figure 1. Clustering of toponyms derived from proper nouns

2.1.1. Extraction of Proper Noun Derivatives

In this paper, the extraction problem of proper noun derivatives is transformed into a mask prediction problem, and the Bert pre-trained language model is selected to predict proper noun derivatives by using the method of prompt learning. The model has good performance in the field of information extraction [11,12]. In ordinary toponyms, toponym proper noun have the function of distinguishing similar geographical entities. Similar geographical entities usually have unique toponym proper noun, and the category of geographical entities has a weak co-occurrence relationship with the toponym proper noun. However, in the set of toponyms derived from the same toponym, the toponyms of similar geographical entities contain the same proper noun derivatives, and the proper noun derivatives in the toponym have a strong co-occurrence relationship with the category of geographical entities. Therefore, this paper uses the toponym category information of the geographical entity as a cue word, and constructs the model input by the way of toponym category + toponym (as shown in Table 1), so that the Bert language model can learn the association information between the proper noun derivatives of the derived toponym and the toponym category, so as to use the model to extract the toponym derivative sequence from the toponym.

 doi:10.20944/preprints202404.0941.v1

Table 1. Input construction of proper noun derivatives extraction model.

Input	The category of toponym is X. The toponym is Y. The proper noun derivatives of toponym is [MASK].	
Example Type	Postive	Negtive
Input[X]	Lake Grove Club	Slichter Residence Hall
Input[Y]	Residential	Building
Output	Lake	None
Answer Map	Lake	Nothing

2.1.2. Extraction of Proper Noun Derivatives

In terms of the formation form of place name words, the place name derived by direct derivation has the same form as its original place name, such as "Columbia" in the place name "Columbia Generating Station" and "Columbia" in its original place name "Columbia River"; while the place name derived by indirect derivation has high similarity to the place name part of its original place name. Such as "Yorker" in the place name "New Yorker Avenue" and "York" in its original place name "New York County". In this regard, this paper selects the word frequency and word character length of the characters that make up the place name words as the word morphological characteristics of the toponym derivation part. For example, the morphological characteristics of "Yorker" are $\{"characterfeature" : "Y" : 1, "o" : 1, "r" : 2, "k" : 1, "e" : 1, "lenfeature" : 6\}$, in which the character features are vectorized by the bag of words model; and then the toponym derivations are clustered by the clustering algorithm, so as to realize the recognition of each toponym derivation set.

2.2. Discrimination of Derivative Relation of Proper Noun

The derived toponyms derived from the same original place name have high similarity. After clustering by clustering algorithm, the toponyms in each derived place name set contain similar toponym derived parts, and it is still impossible to determine which place name is the original toponym and which place name is the derived toponym. In response to this problem, this paper first uses the method of full permutation enumeration to enumerate all possible toponym derivation relationships; then uses the XGBoost model classifier to discriminate various possible toponym derivation relationships, and identifies the derived toponyms and their native toponyms from the set of toponym derivation relationships (as shown in the Figure 2 below). Compared with other classifiers, XGBoost iteratively optimizes the model and performs better in prediction accuracy. Therefore, this paper uses the XGBoost classifier as the discrimination model for toponym derivation relationships[13, 14].

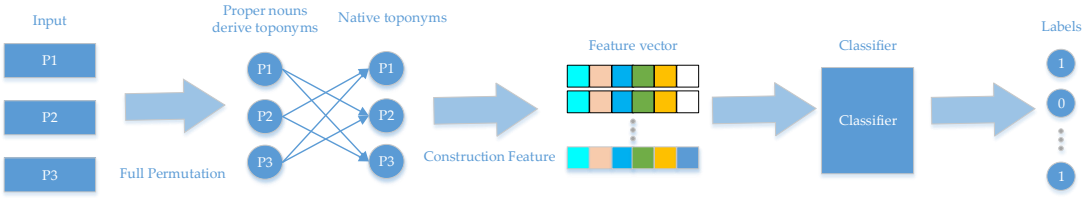

Figure 2. Clustering of toponyms derived from proper nouns

2.2.1. Feature Construction of Toponym Derivation

From a geographical perspective, geographical entities with similar toponyms often have strong geographical associations, and under similar geographical and human environments, the way people name features has certain similarities [15,16]. Therefore, the geographical spatiotemporal distribution between native place name entities and derived place name entities often has certain geographical and statistical significance. In this paper, by analyzing and summarizing the characteristics of toponym

 doi:10.20944/preprints202404.0941.v1

derived place name entities and their native place name entities in the topological relationship, spatial metric relationship, place name survival period, place name derivation mode, and place name morphological relationship between toponym derived place name entities and their native place name, as the model features of the special name derived semantic relationship discrimination model based on XGBOOST.

2.2.2. Spatial Topological Relations

Toponyms have positioning functions [17]. In daily life, when describing the spatial location of a geographic entity, people usually choose another geographic entity as a reference, and use the relative position relationship between the reference geographic entity and the target geographic entity to locate. When naming new geographical entities, people often adopt the way of toponym derivation, and select the proper noun in the geographical entity name as the proper noun part of the new geographical entity name, so as to remind people of the relative position of the geographical entity. For example, one can associate from "New York" in "New York University" that the geographic entity is located in "New York City". Semantically, the toponyms derived from proper noun often contain spatial topological relations such as administrative subordination and proximity with the original toponyms. For example, "New Paltz" in "State University of New York at New Paltz" indicates that the place name is administratively subordinate to "New Paltz (town)". In terms of spatial topological relationship," State University of New York at New Paltz "and" New Paltz "exist Within the relationship (as shown in the Figure 3),"Hudson" in "Village of Castleton-on-Hudson" indicates that the place name is located near "Hudson River", and there is a Cross relationship between "Village of Castleton-on-Hudson" and "Hudson River" in the spatial topological relationship (as shown in the Figure 4). From the perspective of spatial topological relationship, there is a specific spatial topological relationship between the derived toponyms of proper noun and the original toponyms. In this paper, the "nine-intersection model" [18] is selected to describe the spatial topological relationship as the feature of the derived toponyms. Features include Contains, Crosses, Equals, Overlaps, Touches, Within, Disjoint, and so on.

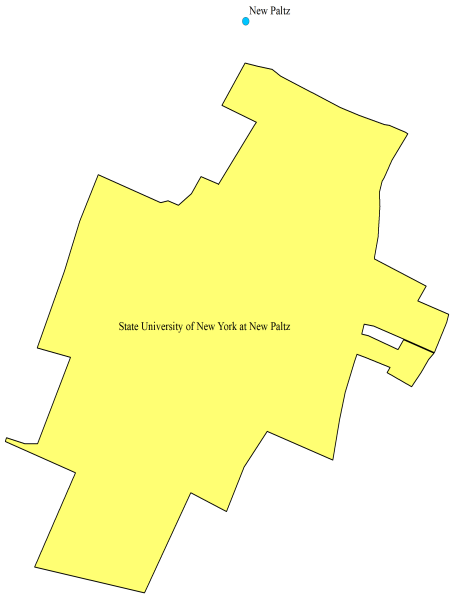

Figure 3. State University of New York at New Paltz and New York City

 doi:10.20944/preprints202404.0941.v1

Figure 4. Hudson River and Village of Castleton-on-Hudson

2.2.3. Spatial Metric Relation

According to the first law of geography[19], "everything is related to other things, but the close things are more closely related", and the association relationship between toponyms is also applicable to the first law of geography [20]. From the perspective of toponyms, in order to express the relative position relationship of the ground object, people usually choose the name of the adjacent geographical entity with high visibility and spatial reference value as the original place name in the process of naming new geographical entities, and name the place by means of toponym derivation. In terms of the form of place name words, the closer the place name is to the original place name, the stronger the correlation of the place name. From the perspective of digitation of toponyms, when the distance between the toponym derived place name entity and the original place name entity is closer, the digitation of toponym derived place name is stronger, and vice versa. In terms of spatial location, the toponym derived place name usually implies the proximity relationship with the original place name. By means of toponym derivation, people can make the place name have the function of "know its name, know its position". For example, people can derive the proper noun "Hudson" from the place name "Hudson Highlands State Park" and know that the place name is located near the "Hudson River" (as shown in the Figure 5). Therefore, this paper chooses the distance between the derived toponym entity and the native toponym entity as the feature of the toponym toponym derivation relation. hudsonriver1

Figure 5. Hudson Highlands State Park

 doi:10.20944/preprints202404.0941.v1

2.2.4. Place Name Lifetime

Toponyms have three basic characteristics of space, attribute and time as common geographical phenomena [21,22]. The starting time and ending time of the spatio-temporal state of a place name is called the life period of the place name. Toponym derivation usually occurs between old toponyms and new toponyms. When naming new geographical entities, people tend to choose the toponyms that have been used for a long time and have a wide range of dissemination in human society and are still in use as the original toponyms. From the perspective of time dimension, the beginning time of the original place name is earlier than the beginning time of the place name derived from the proper name; the end time of the original place name is later than or equal to the end time of the place name derived from the proper name. When the difference between the starting time of the lifetime of the original place name and the starting time of the derived place name is larger, the longer the history of the original place name is, the more likely there is a toponym derived relationship between the two toponyms. Similarly, when the difference between the end of the life time of the original place name and the toponym derived place name is larger, the survival state of the original place name is more stable, and the potential change cost of the toponym derived place name is lower. People are more inclined to choose the place name as the original place name when naming the place name by the toponym derived way. Therefore, in this paper, the difference between the start time (as shown in equation 1) and the end time (as shown in equation 2) of the lifetime of the toponym is selected as the toponym lifetime feature of the toponym derivation relation.

$$dif_s = native_s - derived_s \tag{1}$$

$$dif_e = native_e - derived_e \tag{2}$$

In equations 1 and 2, $native_s$ and $native_e$ are the start time and end time of the native place name lifetime, $derived_s$ and $derived_e$ are the start time and end time of the derived place name lifetime, respectively.

2.2.5. Patterns of Toponym Derivation

Toponym derivation pattern refers to a toponym derivation pattern between one type of native toponym and another type of toponym derivation relationship. In the phenomenon of toponym derivation, the relation of toponym derivation has a direction, and the direction is from the native place name to the toponym derived place name. In the toponym derivation relation, the toponym derivation pattern is mainly divided into: heterogeneous derivation and homogeneous derivation. 1) Heterogeneous derivations are derivations of proper noun between geographical entities with different categories. Native place name entities are usually physical geographical entities and administrative division entities with large spatial span and high visibility, such as mountains, rivers, states, cities, etc. However, the derived toponym entities are usually artificial features with relatively small spatial span and relatively low visibility, such as schools and parks. At the same time, in terms of spatial distribution, the phenomenon of toponym derivation usually exists between various geographical entities with high spatial dependence, such as County and Bus Stop, State and School, River and Footway, etc. 2) Homogeneous derivations are toponym derivations between homogeneous geographical entities. For example, geographic entities with hierarchical relationships, such as trunk versus branch relationships between rivers: "South Fork Shaw Creek" and "Shaw Creek"; And geographic entities that have specific spatial orientation relations, such as the orientation relations between roads: "Frontage Road" and "North Frontage Road". Therefore, this paper selects the native toponym entity category and the derived toponym entity category in the toponym derivation relation as the toponym derivation pattern features.

 doi:10.20944/preprints202404.0941.v1

2.2.6. Morphological Topological Relations of Toponyms

Derived place names are usually composed of their own proper noun part, derived proper noun part and common name part. When toponyms are named by the way of place name derivation, the derived toponyms have different forms of toponyms. For toponym derivation, in the form of place name words, the direct derived toponym derivation is adopted, and the toponym derivation of the derived place name is the same as that of the corresponding original place name. While indirectly derived proper noun derive toponyms, the derivations of their proper noun part are variants of the proper noun of the corresponding native toponyms, with the same word character fragments between them. Therefore, in the topological relation of toponym morphology, the word composition morphology of the derived toponym and its native toponym has a topological intersection relation. The topological relation features of toponyms selected in this paper mainly include: "Disjoint", "Within", "Contain", "Intersection" and other topological relations.

2.3. Construction of the Supervised Dataset

At present, there are few publicly available datasets of geographical entity toponym derivation semantic relationship, and it is difficult to apply to the construction of large-scale geographical entity toponym derivation semantic relationship supervision datasets by means of manual labeling and toponymic literature textual research. In this paper, a rule-based supervised data annotation method for geographical entity toponym derivation relationship is proposed. According to the characteristics of geographical entity toponym derivation relationship, combined with the method of geoscience statistics and the knowledge of manual first verification, the method constructs the toponym derivation relationship discrimination rules for the toponymic set of each derivation mode.

1) **Spatial topology rules**. In the proper noun derivation relationship, the spatial topological feature values between the proper noun-derived place name entity and the original place name entity should be included in the spatial topological feature sequence of the proper noun derivation relationship in the derivation mode. In this paper, the statistical method is used to first count the frequency of the spatial topological relationship values between the place name entities in each derivation mode from the set of place names derived from the proper name, and then the mean value of each spatial topological relationship value is taken as the threshold value, and the spatial topological relationship value whose frequency is greater than the threshold value is added to the spatial topological feature sequence of the derived mode; when the spatial topological relationship values between the proper name-derived place name entity and the native place name entity are included in the spatial topological sequence under the derivation mode, the pair of place name entities satisfies the spatial topological rule.
2) **Spatial measurement rules**. In the toponym derivation relationship, there is a limit to the distance between the derived place name entity and the original place name entity, which is called the derivation distance of the original place name. The distance between the derived place name entity and the original place name entity should be smaller than the derivation distance of the original place name entity. In this paper, the method of Liu Hanyou et al [9] is used to estimate the derivation distance.
3) **place name name lifetime rules**. In the toponym derivation relationship, the lifetime interval of the place name derived from the toponym should be included in the lifetime interval of the original place name. Therefore,Equation 1 and Equation 2 should both be greater than 0.
4) **Toponymic derivation mode rules**. For the $DerivedMode(N \rightarrow D)$of derived toponyms and native toponyms in the toponym derivation relationship, this paper adopts the method of combining artificial prior knowledge to manually select the set of place name derivation modes$D(n_1, s_1), (n_2, s_2), ..., (n_m, s_m)$ that conform to the common sense of place name derivation from the place name category sequence of the set of derived toponyms. In the special noun derived toponyms dataset, the $DerivationMode(s_i \rightarrow n_i)$ composed of the special noun derived place name category and the

 doi:10.20944/preprints202404.0941.v1

original place name category should be included in the artificially defined derivation mode set D.

5) **Topological rules of toponyms**. In the toponym derivation relationship, there is a strong similarity between the derived place name and the toponym part of the original place name. Therefore, the word morphology of the derived place name and the original place name should have a topological intersection relationship. In this paper, the character sequence of the place name word is taken as the morphological feature of the place name, and whether the two toponyms contain the same or similar words is used as the criterion for judging whether there is a topological intersection relationship between the two toponyms.
6) **Construction rules for positive and negative samples**. When both place name entities satisfy the spatial topology rules, spatial metric rules, place name survival rules, place name derivation mode rules, and place name morphological topology rules, then there is a toponym derivation relationship between the two place name entities, and the two place name entities are positive samples; conversely, the two toponyms are negative samples.

3. Results and Discussion

This experiment is mainly divided into three parts: recognition of toponyms derived from proper noun, clustering of toponyms derived from proper noun, and extraction of toponyms derived from proper noun. The experimental data comes from the vector data of 12,733 geographical entities in Illinois, Mississippi, North Carolina, Washington, Wisconsin and other states downloaded from Geofabrik's official website. The experimental objects mainly include the toponym derivation relationship between native place name entities such as State, River, County, and other geographical entities.

3.1. Recognition of Toponyms Derived from Proper Noun

The recognition experiment of proper-name derived toponyms is mainly divided into two parts: the extraction of proper noun derivatives and the clustering of proper noun derivatives. The experimental environment of proper noun derivative extraction is Python 3.9.16, Pytorch 2.0.0 + cu117. The training super parameters of the pre-trained language model are mainly: batch size is 4, learning rate is 2e-5, epoch is 20, and adamW's weight decay is 1e-3. The experiment uses accuracy, precision, recall, and f1 as the evaluation indicators of this experiment (as shown in Table 2). Due to the large number of proper noun derivatives extracted, in order to speed up the training speed of the model, this experiment first divides the geographical entity toponymic data into three parts, then constructs the training dataset, validation set and test set in a ratio of 14:3:3, and finally averages the evaluation indicators of the three test datasets (as shown in Table 3). The clustering of proper noun derivatives uses Homogeneity, Completeness, V-Measure, Adjusted Mutual Information and other indicators as the clustering evaluation indicators of the toponym derivative clustering experiment.

Table 2. Evaluation metrics for derived word extraction results.

Model Name	Accuracy	Precision	Recall	F1
$Bert_{prompt1}$	99.32%	99.75%	99.32%	99.43%
$Bert_{prompt0}$	94.98%	97.34%	94.98%	95.63%
$Ernie_{prompt1}$	97.90%	98.72%	97.90%	98.15%
$Ernie_{prompt0}$	91.54%	92.58%	91.55%	91.30%
$Robert_{prompt1}$	98.79%	99.14%	98.79%	98.85%
$Robert_{prompt0}$	98.36%	97.94%	98.36%	98.10%

* The subscript prompt1 indicates that the promot template for the place name category has been added, and the subscript prompt0 indicates that the prompt template for the place name category has not been added.

 doi:10.20944/preprints202404.0941.v1

Table 3. Evaluation index of toponymic clustering results derived from proper noun

Model Name	Homogeneity	Completeness	V-measure	Adjusted Mutual Information
DBSCAN	84.19%	82.17%	83.17%	82.94%
GaussianMixture	84.37%	81.08%	82.69%	82.47%
Kmeans	81.50%	77.52%	79.46%	79.20%
Kmeans++	81.68%	78.25%	79.93%	79.67%

In the extraction experiments of proper noun derivatives, it is shown in several sets of comparative experiments that the extraction effect of proper noun derivatives is improved by adding the category of the place name as a cue word. The experiments show that the toponym derivative extraction method based on the PromptBert model can effectively extract proper noun derivatives from the place name data, and Bert has higher performance than the Ernie and Robert models in the extraction of proper noun derivatives of toponyms. At the same time, in the clustering experiment of proper noun derivatives, it is shown that by constructing the morphological characteristics of proper noun derivatives, the derived toponyms in various clustering algorithms have achieved higher clustering effect. The experiment shows that the characteristics of proper noun derivatives are selected rationally, and it shows that DBSCAN has better clustering effect than GaussianMixture, Kmeans, Kmeans ++ in terms of clustering of proper noun derivatives.

3.2. Discrimination of Toponym Derivation

In the extraction experiment of the toponym derivation relationship, the indicators such as accuracy, precision, recall, and f1 were used as the evaluation indicators of the toponym derivation relationship discrimination experiment (as shown in Table 5). Due to the lack of public datasets of toponym derivation and its native toponyms, this experiment constructs the label dataset through artificially defined discriminant rules; at the same time, since there are relatively few correct positive examples in various possible toponym derivation relationships, this experiment adopts a down-sampling method to construct the positive dataset and the negative dataset 1:1.

Table 4. Evaluation metrics for derived word extraction results.

Rule Type	Rule
Positive example	1)The spatial distance between geographical entities whose derivation mode is $DerivedMode(river \rightarrow *)$ should be less than 2554761.8275.2)The distance between geographical entities whose derivation mode is $DerivedMode(county \rightarrow *)$ or $DerivedMode(state \rightarrow *)$ should be equal to 0. 2)The spatial topological relation sequence of the original geographical names of County and State derived from their proper names is [Contains, Crosses].2)The spatial topological relationship sequence between the original place names of River and their derived names is [Disjoint]. 3)The difference between the beginning time and the end time of the survival period of the original place name and the derived place name is greater than 0. 4)Topological morphology relationship of place names of original and derived place names is Intersection. 5)The artificially defined specific name derivation modes are $DerivedMode(river \rightarrow *; county \rightarrow *; state \rightarrow *)$.
Negative example	If any of the positive example rules is not met, it is a negative example.

.

 doi:10.20944/preprints202404.0941.v1

Table 5. Evaluation metrics for derived word extraction results.

Model Name	Accuracy	Precision	Recall	F1
Xgboost	99.32%	99.33%	99.11%	99.55%
RandomForest	98.07%	98.00%	98.21%	98.10%
GaussianNB	81.34%	84.30%	77.90%	80.97%

In the discrimination experiment of the toponym derivation relationship of geographical entities, it is shown that by constructing the features of the place name semantics, time and space distribution of the toponym derivation relationship, good relationship discrimination results can be achieved in various classification models. The experiment shows that the method extracted in this paper based on Xgboost has good results, which reflects the rationality of the selection of the characteristics of the toponym derivation relationship(In this paper, the weight of features in the model is used to represent the importance index of features as shown in the Figure 6), and shows that in the field of toponym derivation relationship discrimination, Xgboost has better performance than RandomForest and GaussianNB models.

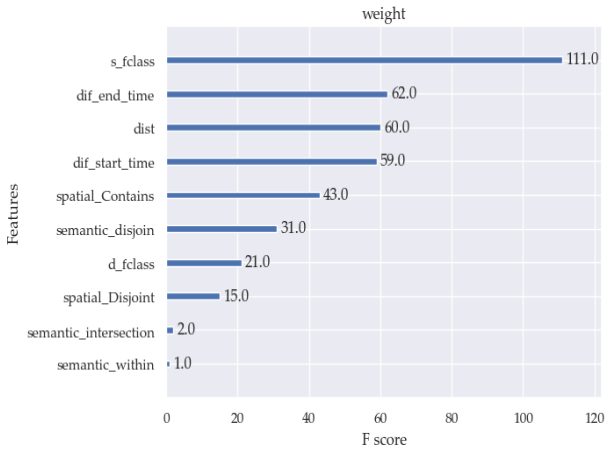

Figure 6. Proper noun derivation relation feature importance evaluation index

4. Conclusions

Toponym derivation is a common phenomenon of place name derivation in geographic entity data. Different from the traditional method of extracting the semantic relationship of geographical entities from text data, the toponym derivation semantic relationship cannot be extracted from the place name data only from the semantic features between geographical entities. This paper combines the toponymic features of native and derived toponyms and the geographical spatiotemporal relationship characteristics between geographical entities in the toponym derivation relationship, and uses the methods of prompt learning, unsupervised clustering and supervised classification to identify the toponym derivation semantic relationship between place name entities from the place name data and its geographic entity vector data, and achieves good experimental results. The semantic relationship recognition method of geographical entity toponym derivation proposed in this paper fills the gap in the extraction of geographical name toponym derivation relationship. This method effectively solves the extraction problem of special name derivation relationship of large-scale geographical entities. However, this method relies too much on artificially constructed supervised data, which requires artificial combination of prior knowledge to construct the characteristics of various special name derivation patterns. In the future, it is necessary to conduct unsupervised mining research on special name derivation patterns.

Author Contributions: Conceptualization,Liu Hanyou; methodology, Liu Hanyou; software,Liu Hanyou; validation, Liu Hanyou; formal analysis, Liu Hanyou; investigation, Liu Hanyou; resources, Liu Hanyou; data curation,

 doi:10.20944/preprints202404.0941.v1

Liu Hanyou; writing—original draft preparation, Liu Hanyou; writing—review and editing, Liu Hanyou; visualization,Liu Hanyou; supervision,Wang Jizhou; project administration,Wang Jizhou; funding acquisition,Wang Jizhou. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: Data will be made available on request.

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

References

1. Yuan, S.; Yang, D.; Liu, J.; Tian, S.; Liang, J.; Xiao, Y.; Xie, R. Causality-aware Concept Extraction based on Knowledge-guided Prompting **2023**. [arXiv:cs.CL/2305.01876].
2. He, Q.; Chen, G.; Song, W.; Zhang, P. Prompt-Based Word-Level Information Injection BERT for Chinese Named Entity Recognition. *Applied Sciences* **2023**, *13*. https://doi.org/10.3390/app13053331.
3. Rozhkov, I.S.; Loukachevitch, N.V. Prompts in Few-Shot Named Entity Recognition. *Pattern Recognition and Image Analysis* **2023**, *33*, 122–131. https://doi.org/10.1134/S1054661823020104.
4. Ashok, D.; Lipton, Z.C. PromptNER: Prompting For Named Entity Recognition **2023**. [arXiv:cs.CL/2305.15444].
5. Chen, P.; Wang, J.; Lin, H.; Zhao, D.; Yang, Z. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. *Bioinformatics* **2023**, *39*, btad496. https://doi.org/10.1093/bioinformatics/btad496.
6. Cheng, H. Derived Place Names and Their Translation Methods. *Chinese scientific and technological terms* **2016**, *18*, 5–8. https://doi.org/10.3969/j.issn.1673-8578.2016.04.001.
7. Ruwei, T. Native place names and derived place names. *Chinese place name* **2018**, p. 30.
8. Cosimo, P. THE CITIES OF THE AMERICAS IN MODERN TIMES: A CASE STUDY ON TOPONYMY. *GEOGRAPHY* **2016**, *9*, 28–46. https://doi.org/10.15356/2071-9388_01v09_2016_03.
9. Hanyou, L.; Jizhou, W.; Xi, M.; Weijun, M. Mining of Complete Derived Place Names and Generic Derived Place Names. *Surveying and mapping science* **2022**, *47*, 176–181+220. https://doi.org/10.16251/j.cnki.1009-2307.2022.10.023.
10. Hanyou, L. Automatic Recognition and Translation of English Derived Place Names for Global Mapping. Master's thesis, Liaoning Technical University, 2022. https://doi.org/10.27210/d.cnki.glnju.2022.000763.
11. Shi, P.; Lin, J. Simple BERT Models for Relation Extraction and Semantic Role Labeling, 2019, [arXiv:cs.CL/1904.05255].
12. Zhao, L.; Li, L.; Zheng, X. A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts, 2020, [arXiv:cs.CL/2001.05326].
13. Zhandong, W.; Shuqin, L.; Sheng, L.; Xinzhi, S. Multi-Class Disturbance Events Recognition Based on EMD and XGBoost in ϕ-OTDR. *IEEE ACCESS* **2020**, *8*. https://doi.org/10.1109/ACCESS.2020.2984022.
14. Li, S.; Zhang, X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing and Applications* **2020**, *32*, 1971–1979. https://doi.org/10.1007/s00521-019-04378-4.
15. Axing, Z.; Guinian, L.; Chenghu, Z.; Chengzhi, Q. Geographic Similarity: Third Law of Geography? *Journal of Geoinformation Science* **2020**, *22*, 673–679. https://doi.org/10.12082/dqxxkx.2020.200069.
16. Li-li, S. A study on the linguistic dimensions of mountain names in China and America. *Journal of Shangluo University* **2021**, *35*, 48–54. https://doi.org/10.13440/j.slxy.1674-0033.2021.01.010.
17. Jian, C.; Bin, Z.; Rupeng, L. Place Name: The "Locator" of Map Spatial Cognition. *Geographic Information World* **2012**, *10*, 6–9+13.
18. Wu Jianxin, Fang Yu, C.B. Research status and development of topological spatial relation description theory. *Geography and Geographic Information Science* **2005**, pp. 1–4.
19. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* **1970**, *46*, 234–240.
20. Xingguang, W.; Ruijie, Z.; Yi, Z. Research on place name disambiguation method based on geographical association degree and evidence theory. *Journal of Peking University (Natural Science Edition)* **2017**, *53*, 344–352. https://doi.org/10.13209/j.0479-8023.2016.090.

 doi:10.20944/preprints202404.0941.v1

21. Jiangtao, B.; Wei, P.; Yongjian, H.; Yongqiang, Z.; Huihui, Y. Research on the Construction of Historical Place Name Comprehensive Information System Based on TGIS and Big Data Technology. *Journal of Global Change Data* **2021**, *5*, 363–372+520–529.
22. Hui, D. Spatio-temporal attributes and cultural values of Chinese place names. *Civil Administration of China* **2017**, p. 55.