International Journal of
Geo-Information

Article

An English-Chinese Machine Translation and Evaluation Method for Geographical Names

Hongkai Ren [1,2,*], Xi Mao [1], Weijun Ma [1,3], Jizhou Wang [1] and Linyun Wang [2]

[1] Chinese Academy of Surveying & Mapping, Beijing 100830, China; maoxi@casm.ac.cn (X.M.); weijunma@casm.ac.cn (W.M.); Wangjz@casm.ac.cn (J.W.)

[2] College of Geomatics, Shandong University of Science and Technology, Qingdao 266590, China; linyun330@foxmail.com

[3] Capital Normal University, Beijing 100048, China

* Correspondence: sparenopains@foxmail.com; Tel.: +86-18811339420

Received: 22 January 2020; Accepted: 20 February 2020; Published: 25 February 2020

Abstract: In recent years, with increasing international communication and cooperation, the consensus of toponymic information among different countries has become increasingly important. A large number of English geographical names are in urgent need of translation into Chinese, but there are few studies on machine translation of geographical names at present. Therefore, this paper proposes a method of automatically translating English geographical names into Chinese. First, the lexical structure of the geographic names is analyzed to divide the whole name into two parts, the special name and the general name, in an approach based on the statistical template model that implements pointwise mutual information and a directed acyclic graph data structure on the extracted names from different categories of a geographical name corpus. Second, the two parts of the geographic names are translated. The general name can be directly translated via methods of free translation. For the transliteration of the special name, the phonetic symbols are generated based on the cyclic neural network, and then, the syllables are divided based on the minimum entropy and converted into Chinese characters. Finally, the two parts of Chinese characters are combined, and criteria are prepared to evaluate the translation reliability according to the translation process to realize automatic quality inspection and screening of geographical names. As the experimental results show, the method is effective in the translation process of English geographic names into Chinese. This method can be easily extended to other languages such as Arabic.

Keywords: machine translation; phonetic symbols generation; syllable division; cyclic neural network; minimum entropy; lexical structure analysis; automatic evaluation

1. Introduction

The geographical name [1] is a special name given to a geographical entity [2] in a specific spatial location and is also an essential geographic information element in the spatial database. The global strategic deployment of "the Belt and Toad" urgently requires support from the geographical name. However, most of the overseas geographical names do not have a Chinese expression. Geographical name databases contain only a few small-scale data [3], which cannot meet the increasing demands of the construction of global geographic name information resources. At present, geographical names are mainly translated manually, with some shortcomings such as low efficiency, high cost and difficulty in verifying errors in the context of large-scale operations. As English is the most widely used language in the world, determining how to achieve efficient and accurate translation of English geographical names is particularly important for enriching global geographic information resources.

At the same time, we have known that it is possible to translate English geographical names efficiently because of the development of machine translation [4]. However, geographical names are composed of a group of phrases without clear grammatical constraints, so the lexical structures at different scales are complex and diverse. Second, the translation of geographical names needs to follow relevant standards [5]. For example, a geographical name is composed of 2 components: the special name and the general name [6]. The general name is a word or phrase that describes the category of a geographical entity, while the special name is a special word used to distinguish different individuals in a certain kind of geographical object. For example, in a simple English geographical name, such as "Glenore Railway Station", usually "Glenore" is recognized as a special name and "Railway Station" is the general name. According to national English–Chinese translation guidelines, transliteration of the special name and free translation of the general name should be guaranteed, which ensures the accuracy and applicability of geographical names over a wide range. Correspondingly, establishing how to correctly distinguish between the special name and the general name of a geographical name is very crucial [7]. Therefore, general translation models such as Google translation [8,9], which particularly emphasize free translation, are not applicable to the translation of geographical names. At present, there are few studies worldwide that address automatic translation of geographical names from English to Chinese [10,11]. Most of them focus on transliteration [12], which is based on simple table matching or manual input to obtain phonetic symbols, and the syllable segmentation approach has defects of intersection ambiguity and low accuracy. The related research does not consider the category attribute of the geographic entity and does not solve the key problem of the reasonable distinction between the special name and the general name of the geographical name. In addition, the final step in the production of geographical names requires the quality inspection of the translation results. At present, the quality inspection is purely manual, which is time consuming and laborious.

To solve the above problems, this paper establishes an English–Chinese machine translation and evaluation model for geographical names based on the word-formation characteristics and attribute information of English geographical names, the theoretical knowledge related to machine learning [13] and the geographical name translation standards of China [14]. The basic thinking is as follows: first, all toponymic data are divided into groups according to the category attributes; then, based on pointwise mutual information [15], common phrases of different categories of the corpus are calculated and explored [16], and the data structure of a directed acyclic map [17] is used to extract the geographical name template. In the process of translation, the same category of the template is used to nest matching geographical names and split their structures completely to generate a lexical structure tree [18,19]. This tree contains two parts: the special name and the general name. Machine learning is used to transliterate special names, while general names are translated freely according to corresponding templates. Each part of the tree is converted into Chinese by a cyclic recursive method, and finally combined to complete the translation of the geographical name. Finally, the evaluation index [20,21] is set, and the expression is constructed to simulate the intermediate process of geographical name translation. The reliability of geographical name translation is measured [22] according to the index value, and the automatic quality inspection of geographical name translation is realized.

2. Method

2.1. Construction of the Geographical Name Template based on Pointwise Mutual Information and Directed Acyclic Graph

2.1.1. Pointwise Mutual Information

Mutual information (MI), also known as trans-information, is a measurement that describes interdependence among variables in information theory [23]. Generally, the mutual information of two discrete random variables X and Y can be defined as:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ represents the joint probability distribution function for X and Y, and $p(x)$ and $p(y)$ are marginal probability distributions for X and Y, respectively. Intuitively, pointwise mutual information is used to measure the shared information of X and Y, or given one of two variables, reduce the degree of uncertainty of the other.

Pointwise mutual information (PMI) [24,25] refers to a method to measure the probability of the simultaneous occurrence of two random events in a given joint distribution and edge distribution under the assumption of independence, and mainly focuses on a single probability event compared with mutual information. Its expression is defined as:

$$PMI(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

In view of the expression, mutual information is the expectation of pointwise mutual information. In computer linguistics, pointwise mutual information is applied to search collocations and connections between words. The occurrence probabilities of the two words are taken as approximations of the edge distributions [26] $p(x)$, $p(y)$, and their co-occurrence probabilities are approximated by the joint distribution $p(x,y)$. In the translation of the geographical name, the general name of a geographical name has the characteristics of defining the type of surface features, and the special name of a geographical name has the characteristics of distinguishing the same kind of surface features. Therefore, the general name often shows a fixed word/phrase collocation with a high PMI, while there is no significant correlation among special names. Based on this, a template extraction method of the geographical name based on point mutual information is constructed.

2.1.2. Template Expression of Geographical Name

According to the analysis of the composition of geographical names, the free combination of a special name and common words/phrase is expressed as a structural form of a geographical name. The mapping relationship f is as follows:

$$x(S,W) \rightarrow y \tag{3}$$

In the mapping relationship, S is the special name of a geographical name, and W includes the general name of a geographical name, prepositions, conjunctions, adjectives and quantifiers.

In addition, the string of geographical names cannot fully express the geo-spatial object; however, geographical entities also have various attribute information with category attributes that can be used as an important reference in geographical name translation. Moreover, a complete single geographical name must belong to a certain category.

$$\forall x(y(x)) \rightarrow Type(x) \tag{4}$$

In the logical expression, $y(x)$ represents the string x expressed as a geographical name; $Type(x)$ indicates that the string x has a category attribute. GeoNames [27] classified the categories of geographical name data as shown in Table 1.

The geographical name template [28] is an abstract expression form of the geographical name structure [29]. The custom placeholder ([X], [Y], [Z], [M], and [N]) is used to replace the special name S and combine it with common words/phrases W. An example of a single geographical name template is shown below:

$$\text{[X] Railway Station} \rightarrow (\text{[X] 火车站}, 243, S)$$

In the expression, the placeholder [X] is the special name or other nest template, the number 243 is the statistical frequency of this template in the entire geographical name corpus, and S is its category.

Table 1. Categories and descriptions of the geographical name.

Categories	Description	Detailed Coding (Part)
A	Characteristic of administrative division (nation, state, etc.)	A.ADM1,A.ADM2
H	Hydrological characteristic (stream, lake, etc.)	H.BAY,H.LK
L	Specific functional areas (oilfield, farm, etc.)	L.PRT,L.PRK
P	Population aggregation characteristic (city, village, etc.)	P.PPL,P.PPLQ
R	Traffic characteristic (highways, railways, etc.)	R.RD,R.TNL
S	Building facilities point (hospital, railway station, etc.)	S.HSP,S.CH
T	Geological features (mountains, islands, etc.)	T.MT,T.ISL
V	Vegetation characteristic (forest, wasteland, etc.)	V.GRSLD,V.SCRB
U	Seabed characteristic (shoal, trench, etc.)	U.SHLU,U.RFU

2.1.3. Template Extraction Algorithm

First, all geographical name corpora are classified into nine categories, and then data are trained successively. The specific process entails traversing a certain category of the geographical name corpus to count the word pairs $(Count_{ab})$ and the number of single words $(Count_{a})$ in each geographical name and separately calculate their probabilities. There is a high probability of multiple occurrences of a single word or word pairs, according to the pointwise mutual information formula,

$$PMI_{ab} = \frac{P_{ab}}{P_a P_b} \tag{5}$$

Through the calculation of the value of each word pair, the threshold $e1$ is set, and the word pairs in $PMI_{ab} > e1$ are stored in set H. Then, each geographical name is traversed again, and a directed acyclic graph is established one by one. The words in the sentence act as points on the graph: in the sentence, given the ordered word pairs (a, b), if $(a, b) \in H$, then a directed edge "$a \rightarrow b$" is added to the graph; otherwise, it will not be processed. After the traversal, all edges of the digraph are determined. For each path, the corresponding words of the node are accessed in turn (if the adjacent nodes are crossed, then a placeholder is inserted, such as [X], [Y]). A geographical name template is recorded and added to the current category template base. Finally, the occurrence frequency of all templates is counted and arranged in an orderly manner according to the size, and the threshold is set, which is selected and determined according to the total number of template bases of different categories.

Figure 1 shows the template extraction of a single geographical name "East Ayrshire Community Hospital", and the following template results are obtained:

(1) East [X]→东[X]

(2) [X] Community Hospital→[X]社区医院

(3) [X] Hospital→[X]医院

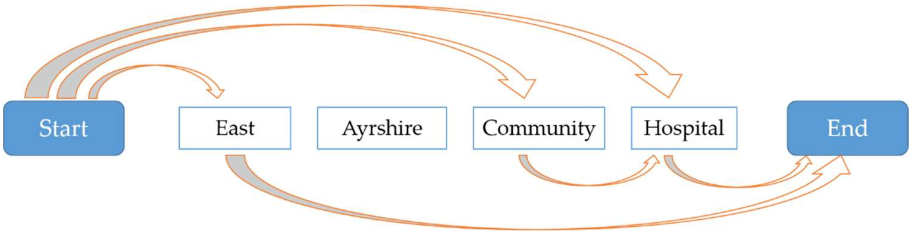

Figure 1. Template extraction diagram of the geographical name.

Different categories of templates to match the corresponding geographical names will ensure the correctness of lexical structure analysis. The creation process of template libraries is shown in Figure 2.

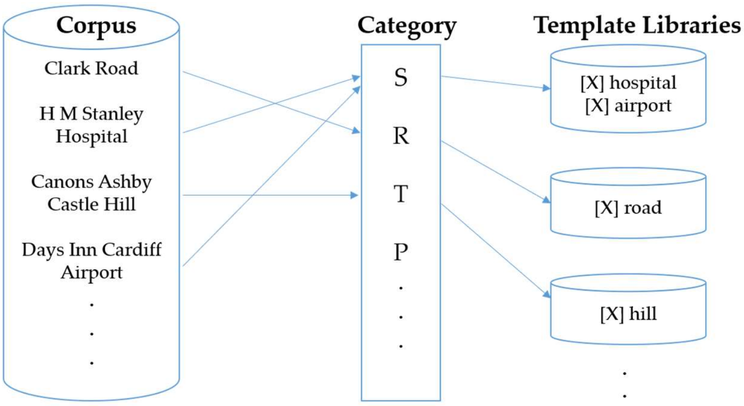

Figure 2. Creating template libraries based on categories of geographical names.

2.2. Analysis of Geographical Name Lexical Structure based on the Statistical Template Model

In natural language processing, syntactic structure analysis [30] refers to processing the input word sequence according to the given grammatical rules to obtain a regular syntactic structure. In the machine translation of geographical names, this paper proposes the following hypothesis based on the "projective hypothesis" of syntactic analysis:

(1) The geographical name is composed of several templates, which do not intersect each other.

(2) Each individual word can be regarded as the simplest geographical name template.

(3) Placeholders of geographical name templates can replace other geographical name templates.

Based on the above assumptions, this paper proposes a decomposition algorithm of the geographical name hierarchical structure [31] as follows: find a series of geographical name templates to cover each word in the geographical name without repetition, omission, or intersections, to maximize the sum of the logarithmic frequencies of these geographical name templates. The flow chart of the algorithm is as follows:

(1) Determine the template base according to the category of the geographical name to be translated.

(2) Scan and determine the nested template combination that complies with the above assumptions to completely split the geographical name and establish the lexical structure tree. The algorithm mainly builds a circular matching function GHSS (G, P), where G is a word storage list of geographical names, and P represents a dictionary of related templates stored in a trie data structure. The scanning algorithm implemented in the Python language environment is shown in Algorithm 1.

Algorithm 1: Lexical structure analysis algorithm for geographical names

```
Input: The geographical name G to be scanned and template set P of the same category
Output: All appropriate decomposition schemes S for the lexical structure of the geographical name
Function GHSS(G as list<str>, P as trie-dict<str, dict>):
for word in G:
if word is in P.keys():
   GHSS(G[1:], P[word])
   record structure
if word is regarded as a placeholder
   GHSS(G[1:], P[placeholder])
   record structure
return structure
```

(3) The logarithmic frequency of nested templates is obtained by summing the logarithmic frequency of each template. The logarithmic frequency of a single template is calculated by the proportion of its occurrence frequency to the total number of occurrences of all templates.

(4) Select the scheme with the maximum probability value and realize translation accordingly.

Given the example of the geographical name "Isle of West Burrafirth", Figure 3 shows the three nested template combination schemes obtained after hierarchical structure analysis (it can be visualized as A, B and C from left to right).

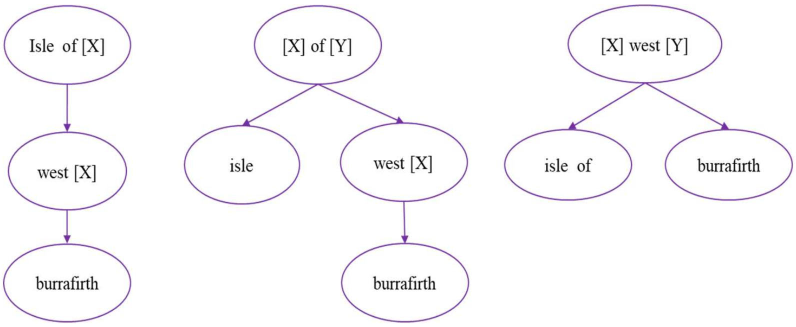

Figure 3. Schemes of lexical structure analysis of the geographical name "Isle of West Burrafirth". All three schemes satisfy established assumptions.

Table 2 shows some templates involved in A, B and C and logarithmic frequencies in the template library of category *T*. For example, the frequency of template "[X] of [Y]" is 148. The total number of occurrences of all templates is 30,221.

Table 2. Templates and logarithmic frequency (partial).

Template	Logarithmic Frequency
Isle of [X]	−7.37185
West [X]	−5.05360
[X] of [Y]	−5.31908
[X] west [Y]	−6.36505

Observation of scheme A shows that it consists of two templates, (Isle of [X] and West [X]), and a single word (Burrafirth), and the logarithmic frequencies of the two templates are recorded as $P1$ and $P2$. The individual word uses the a priori value P_{word}, which is defined as −11.00944 (the logarithmic frequency is calculated by assuming the number of individual word frequencies is 1/2). Then, the logarithmic frequency sum P in the scheme can be expressed as:

$$P = \ P1 + \ P2 + P_{word}$$

Finally, after calculation, scheme A is −23.43489, B is −32.39156 and C is −38.08192. Therefore, scheme A is determined as the final structure tree of the geographical name. The translation process according to the nested templates is as follows:

(1) Isle of [west burrafirth] → [west burrafirth]岛
(2) [west[burrafirth]]岛 → [西[burrafirth]]岛
(2) [西[burrafirth]]岛 → [西[巴勒弗斯]]岛
(4) [西[巴勒弗斯]]岛 → 西巴勒弗斯岛

2.3. *Transliteration of Special Names based on Machine Learning*

The transliteration of special names of the geographical name is actually the problem of solving the mapping between words to phonetic symbols and phonetic symbols to Chinese characters. The transliteration process of the geographical name is shown in Figure 4. From left to right, it can be divided into five steps: geographical name input, phonetic symbols generation, syllable division, syllable conversion to Chinese characters and result output.

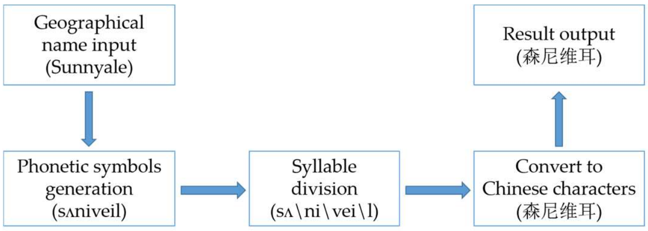
Geographical name input (Sunnyale)
Result output (森尼维耳)
Phonetic symbols generation (sʌniveil)
Syllable division (sʌ\ni\vei\l)
Convert to Chinese characters (森尼维耳)

Figure 4. Automatic transliteration technology framework for geographical names.

2.3.1. Phonetic Symbol Generation Method based on the Cyclic Neural Network

Considering the similarity between the process of generating words to phonetic symbols and the process of generating another language from one language by machine translation, accepting one random sequence and outputting another random sequence is shown by model calculation. Based on the idea of the decoder and encoder [32] in end-to-end neural machine translation [33], this method proposes a phonetic symbol generation model based on the encoder–decoder structure for processing the phonetic symbol generation of words.

The basic network construction of this method is shown in Figure 5. The input character set E is 26 English letters with start and stop symbols <s>, </s>, and the output character set F is 55 phonetic letters with start and stop symbols. Each character is mapped to d-dimensional space and initialized as a real vector $v \in R^d$. This step is to construct embedded characters. Each embedded character is stored in the matrix $M_E \in R^{|E|*d}$ or $M_F \in R^{|F|*d}$ by type and can be searched by id. A concrete example is shown in Figure 4. First, the word "york" is split into individual letters: y, o, r, and k. The letters are then converted into 4 real value vectors by embedding the matrix M_E and are progressively entered into the encoder. The encoder's coding unit, when accepting a letter input, combines the dense vector of the previous state code, encodes it into a new dense vector and passes it to the next state. After confirming that these words have been entered, the dense vector decoding of the words begins, which is the phonetic symbol output. First, the start symbol "<s>" is input and converted into a real value vector by M_F and combined with the decoding unit of the decoder, and then the decoded vector is calculated together with the dense vector encoded by the encoder. Consequently, the phonetic symbol with the highest probability of occurrence is calculated by the *softmax* function, thus obtaining the first phonetic symbol ″j″. This phonetic symbol is then used as the decoder phonetic symbol for the next moment. This process is repeated until all the phonetic symbols have been generated after the end symbol "</s>" is output.

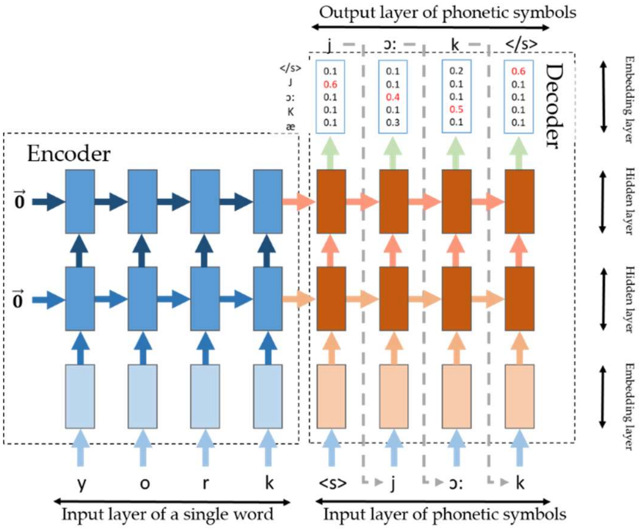

Figure 5. Framework map of the network model for phonetic symbol generation.

In this paper, the decoding unit of the encoder and decoder selects the normalized long short-term memory (LSTM) unit, which has better experimental performance than other cyclic neural networks [34].

Furthermore, the length of the word embedding vector in the actual construction of the network should be equal to the cyclic neural network unit in the hidden layer, which is called the number of hidden layer units.

2.3.2. Syllable Division Method based on Minimum Entropy

The method performs unsupervised learning on the phonetic corpus based on minimum entropy and obtains the syllable probability distribution from the original corpus. According to the distribution, the entropy value of different syllable division modes in the phonetic symbol string is calculated, and the minimum entropy value is selected to uniquely determine the syllable segmentation result.

Entropy is a measure of the amount of information contained in a variable. For a random variable x, its probability density function is P, and its entropy is expressed as $E = -\int P(x)logP(x)$. In discrete cases, it can also be expressed as $E = -\sum P(x)logP(x)$. The principle of minimum entropy is to minimize the entropy E by modifying the probability density function and to achieve the minimum redundant information contained in the variable [35].

This method defines the average entropy of the phonetic symbols $E = -\sum\limits_{x\in\Gamma} P(x)logP(x)$, where $P(x)$ represents the frequency at which the syllable x appears in the corpus, and Γ represents the set of syllables. Then, when some phonetic symbols are combined into a syllable, the average entropy of the phonetic symbols can be expressed as:

$$E = -\frac{\sum_{x\in\Theta} P(x)logP(x)}{\sum_{x\in\Theta} P(x) * l_x} \tag{6}$$

In the expression, *P(x)* represents the frequency of syllable x in the corpus, l_x represents the length of the syllable (determined by the number of phonetic symbols), and Θ represents the set of syllables. The purpose of this method is to find a syllable segmentation set that minimizes E. Intuitively, finding the right syllable set directly is difficult. However, using the GIS (Generalized iterative scaling) algorithm can obtain an approximate solution of the target set [36]. By combining two randomly selected elements a and b in set Θ, a new syllable set $\Theta' = \Theta \cup \{ab\}$ can be obtained. Because the set Θ' is different from Θ, the probability distribution P will also change. Assuming that the probability distribution of the syllable set Θ' is $\hat{P}$, then there is

$$\hat{P}(ab) = \frac{N_{ab}}{N - N_{ab}} = \frac{P(ab)}{1 - P(ab)} \tag{7}$$

$$\hat{P}(i) = \frac{N_i}{N - N_{ab}} = \frac{P(i)}{1 - P(ab)}, \ i \neq a, b \tag{8}$$

$$\hat{P}(i) = \frac{N_i - N_{ab}}{N - N_{ab}} = \frac{P(i)}{1 - P(ab)}, \ i = a, b \tag{9}$$

According to the new probability distribution $\hat{P}$, the entropy E_{ab} after merging a and b can be calculated by Formula (12).

By subtracting E from E_{ab}, the change value of entropy E_Δ after merging a and b is obtained.

$$E_\Delta = E_{ab} - E \tag{10}$$

The expression of E_Δ after calculation is:

$$E_\Delta = -\frac{\Delta}{\sum_x P(x)l_x} \tag{11}$$

where Δ can be expressed as:

$$\Delta = P(ab)log\frac{P(ab)}{P(a)P(b)} - P(ab))log(1 - P(ab))* \sum_{i=a,b}(P(i) - P(ab))log\left(1 - \frac{P(ab)}{P(i)}\right) \quad (12)$$

Using the idea of a greedy algorithm, the last syllable set and the probability distribution can be determined by the two elements that make the entropy decrease the most with each merge. It is also noted that in the case of a fixed corpus C, a syllable set Θ, and its probability distribution P, $\forall a, b \in \Theta$, the denominator of Formula (11) is always constant. A set of a, b is identified to maximize the value of the molecule of Formula (11). Thus, the specific algorithm flow is shown in Algorithm 2:

Algorithm 2: Average minimum entropy for phonetic symbols

```
Input: Phonetic symbols set $\Gamma$, Corpus C
Output: Syllable set $\Theta$, Average minimum entropy for phonetic symbols E, Probability distribution P
function AME($\Gamma$, C)
  $\Theta \leftarrow \Gamma$
  while TRUE do
    a, b $\leftarrow$ argmax$\Delta$(a, b),    a, b$\in\Theta$
    if $\delta$(a, b) < 0 then
      return $\Theta$, E, P
    end if
    E $\leftarrow$ E(a, b)
    $\Theta \leftarrow \Theta \cup \{ab\}$
end while
end function
```

For example, the combination of the phonetic symbols dz and u in syllable set Θ maximizes the molecular value of Formula (11), thereby updating the set Θ (Θ ∪ {dzu}) and the average entropy E of the current phonetic symbols and the syllable distribution probability P. The updating will stop until the syllable set Θ cannot find a set of phonetic symbols to make the molecular value of Formula (11) greater than zero.

In this paper, we use the shortest path algorithm and word segmentation method to obtain several candidate segmentation schemes Ψ, and then select the lowest entropy of several syllable sets as the final segmentation result $\psi = argmin_{\psi \in \Psi} - \sum_{x \in \psi} logP(x)$. For example, for the string "sniveil" of phonetic symbols, the candidate segmentation schemes are "s\ni\vei\l" and "sn\i\vei\l", which are recorded as ψ_1 and ψ_2, respectively. The frequency of the syllables is s: 0.015, ni: 0.02, sn: 0.003, i: 0.01, vei: 0.03, l: 0.02. The entropy of ψ_1 and ψ_2 is 6.74 bits and 7.74 bits, respectively. Therefore, "s\ni\vei\l" is chosen as the final segmentation result. These syllables can be mapped into Chinese characters by the English–Chinese translation table: "s–森", "ni–尼", "vei–韦", and "l–耳". Finally, these Chinese characters are combined in sequence as the result of transliteration of special names.

2.4. Automatic Evaluation of Geographical Name Translation Results

To ensure the translation quality, it requires manually reviewing the translation results after completing the translation of geographical names. In the context of a large-scale operation, manual machine calibration has defects such as low accuracy and high time consumption. This paper takes the process of machine translation of English geographical names as a reference and evaluates the availability of translation results by setting evaluation indexes, which are positively correlated with the reliability of translation.

The translation process of geographical names is composed of transliteration and free translation. The reliability of transliteration and templates determines the accuracy of translation results. Therefore, this paper first calculated the estimated value of the two factors separately, and then calculated the final evaluation index based on the weight of the number of words of the two factors. The formulas are as follows:

$$Acc = \frac{acc_{template} * (Sent_{len} - word_{num}) + acc_{translit} * word_{num}}{Sent_{len}} \tag{13}$$

In Equation (13), $acc_{translit}$ is the transliteration reliability, $acc_{template}$ is the template reliability, $Sent_{len}$ is the total number of geographical names, and $word_{num}$ is the number of special names for transliteration.

The evaluation index calculation complies with the following cognitive thought:

(1) In the process of lexical structure analysis, if the geographical names do not match any template, the default is that all the words are proper names. The index value depends entirely on the reliability of transliteration. In contrast, the influence of template reliability is enhanced. This paper presents a dynamic weighting evaluation method.

(2) The determined special name nouns may contain the customary words already existing in the geographical name database of China. The more these words are used directly in translation, the higher the reliability of transliteration.

(3) The deeper the toponymic lexical structure tree is, the more matched templates there are, the more unstable the lexical structure tree is, and the lower the reliability of the template.

(4) The higher the sum of nested template frequencies, the more reasonable the structural fractal formula, and the higher the reliability of the template.

$$acc_{translit} = BaseAcc + (1 - BaseAcc) * \frac{instan_{num}}{word_{num}} \tag{14}$$

$$acc_{template} = penalty_{template} * RawAcc_{template} \tag{15}$$

$$penalty_{template} = \begin{cases} \exp\left\{1 - \frac{template_{num}}{Para_{num}}\right\}, \ template_{num} > Para_{num} \\ 1, \ template_{num} \leq Para_{num} \end{cases} \tag{16}$$

$instan_{num}$ is the number of special names in the existing database of geographical names. The higher its proportion in all special names, the higher the value of $acc_{translit}$. The rest of the special names are transliterated by the transliteration model of this article. Considering its accuracy, the reference accuracy parameter *BaseAcc* is set as 0.9; $penalty_{template}$ is a penalty item; $template_{num}$ is the total number of matched templates; $RawAcc_{template}$ is the criterion accuracy of the template; and the above parameters are obtained in the process of index calculation. T number of templates matched to all the toponymic experimental data are recorded, and the threshold parameter $Para_{num}$ is set as 3 by calculating its average and variance. $RawAcc_{template}$ is realized by constructing quasi-convex functions. This paper proposes two design ideas: *Sigmod* function implementations and *square* function implementations. The calculation formulas are as follows.

Sigmod function implementations:

$$RawAcc_{template} = sigmod\left(\exp\left(\frac{SentTree_{Score}}{template_{num}} - maxlogp\right) * 1000 - para_{sigmod}\right) \tag{17}$$

Square function implementations:

$$RawAcc_{template} = \exp\left(\frac{SentTree_{Score}}{template_{num}} - maxlogp\right)^{\frac{1}{para_{sqr}}} \tag{18}$$

In the expressions, $SentTree_{Score}$ is the sum of the template frequencies during the decomposition of geographical name structures, *maxlogp* is the maximum template frequency value in all geographical name analyses, while $para_{sigmod}$ and $para_{sqr}$ are hyperparameters. To meet the value range required by the function, a is used as the adjustment parameter and is manually set to 5.

After determining the hyperparameters, the evaluation indexes of all geographical names are calculated, the thresholds are confirmed according to the distribution of the evaluation indexes, and the translation results with high or low reliability are screened.

3. Experiments

3.1. Experimental Data

This paper selects the data of British geographical names downloaded from GeoNames as experimental data, with a total of 62,878 items. The complete attribute information of a single geographical name entry is shown in Table 3.

Table 3. Main fields and descriptions of experimental data.

Field	Description
GeoNames ID	ID in GeoNames database
Name	Geographical names after Romanization
Latitude	Location latitude
Longitude	Location longitude
Feature code	Category coding of location features
Country code	Country code of location
Adamin code	Administrative division-level coding of location
Time zone	Location time zone
Modification data	Current information of last revision time

The distribution of experimental data of geographical names of different categories is shown in Figure 6.

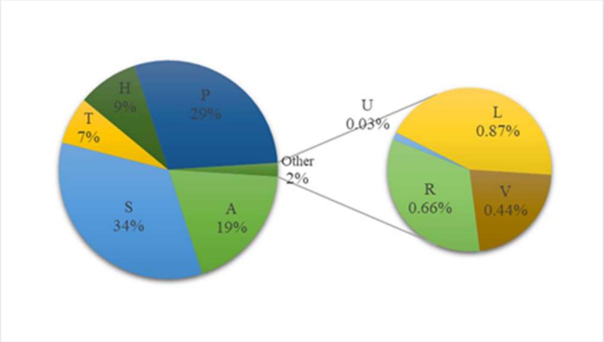

Figure 6. Geographical names in different categories in Britain.

3.2. Experimental Process

The machine translation of English geographical names mainly includes the following steps: geographical names input, template extraction by category, lexical structure analysis, transliteration of special names and automatic evaluation. The whole technical process is shown in Figure 7.

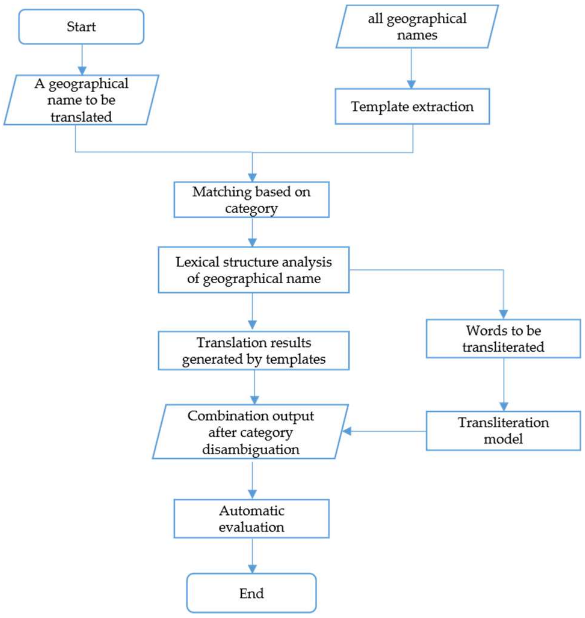
Start
all geographical names
A geographical name to be translated
Template extraction
Matching based on category
Lexical structure analysis of geographical name
Words to be transliterated
Translation results generated by templates
Combination output after category disambiguation
Transliteration model
Automatic evaluation
End

Figure 7. English geographical name machine translation and evaluation flow chart.

3.3. *Experimental Results and Analysis*

3.3.1. Machine Translation

In the experiment of place name translation, 29,686 words were counted, and their occurrence frequency was 13,102 times. In addition, 72,565 pairs of words were counted, and their occurrence frequency was 116,217 times. The probability distribution of words, word pairs and point mutual information is calculated, as shown in Figure 8.

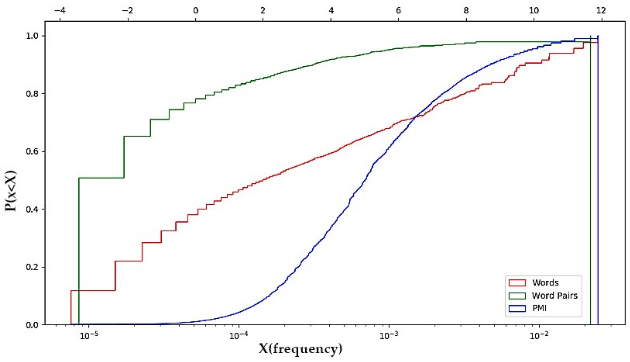

Figure 8. Accumulated distribution map of word frequency, word pair frequency and point mutual information probability.

In the figure, the word frequency (red) and word pair frequency (green) correspond to the lower coordinate axis, with a value range of approximately (8.6*10^{-6}, 0.025), and the point mutual information (blue) corresponds to the upper coordinate axis, with a value range of approximately (−4, 12).

Figure 9 is the corresponding scatterplot of word-pair frequency and pointwise mutual information. The bottom parallel line represents the lower frequency distribution of word pairs. The right small plot shows the word pair with a higher pointwise mutual information value. From top to bottom, from left to right, the order is railway (railway, station), (the, hotel), (holiday, inn), (best, western), (western, hotel), (underground, station) and (guest, house).

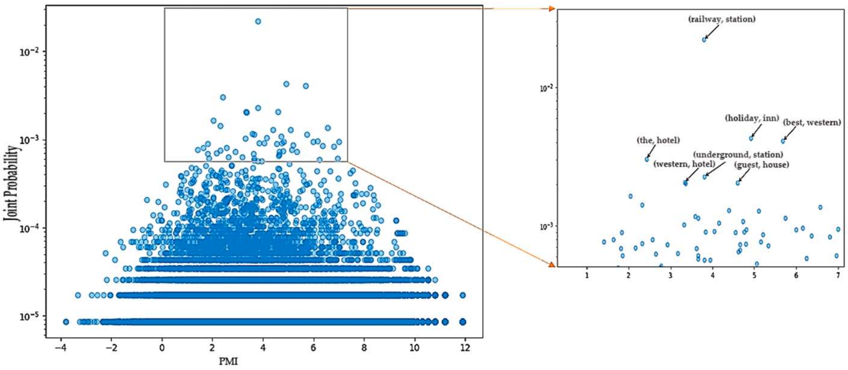

Figure 9. Scatterplot of word-pair frequency and value of pointwise mutual information.

The thresholds of the joint distribution (word-pair frequency) and point mutual information are reasonably set according to Figures 8 and 9. Because the word pairs with a lower value of the joint distribution may also obtain a higher point mutual information value, the word pairs with a frequency of 1, 2, 3, and 4 are discarded in order to achieve a high-quality template, and the threshold of the joint distribution is set to 5*10^{-5}. The point mutual information threshold is set to 3.1, and the first 80% of the word pairs are selected. Similarly, according to the probability accumulation map, the extraction threshold of the template is set to 3, and a total of 876 geographical name templates are obtained, of which 279 belong to the commonly used building facilities (S), which has the highest proportion. Partial templates are shown in Table 4.

Table 4. Results of template extraction of geographical name (partial).

Template of Geographical Names	Frequency
[X] railway station	2059
[X] castle	988
[X] bay	717
[X] and [Y]	510
river [X]	399
[X] of [Y]	148
[X] in the [Y]	48
our lady [X] church	14

In the process of geographical name translation, 4219 place names with the number of words of 3 or more were structurally decomposed, and the average depth of the established trees was 1.8, while the average number of decomposed templates was 2.7. Given that determining whether the template matched by randomly selecting the data of 100 geographical names is reasonably evaluated manually, the acceptable rate is 91%.

Six thousand results of place name translations in this experiment have been randomly selected and sent to a professional translation company for proofreading, with an 84.5% acceptability rate. Meanwhile, 500 geographical names are randomly translated by Baidu Translation and Google Translation, and after manual review, the acceptability rates of translation results are 74.4% and 52.6%, respectively.

The corpus data of the "word-phonetic symbol" used for training comes from the open source English dictionary, with a total of 97,857, and is divided into a training set, verification set and test set according to the ratio of 70:20:10. The training environment of the phonetic symbol generation model is as follows: GPU: GTX970M, CPU: i7-6700HQ, and OS: Windows 10, version 1803. Since the network structure of the phonetic symbol generation model will significantly affect the quality of the results, this paper uses the 10-fold cross-validation method to train different neural network structures. The following are the hyperparameter settings and evaluation indicators of several major network structures. To ensure the quality of the generated phonetic symbols, this paper uses the general accuracy rate, BLEU and ROUGE to evaluate the model. The comparison results are shown in Table 5.

Table 5. Model performance comparison table of different network structures.

Network Structures	Hidden Layers	Units Number	Accuracy Rate	BLEU	ROUGE
Structure 1	2	128	83.3%	92.97%	97.2%
Structure 2	1	128	80.7%	92.2%	96.8%
Structure 5	3	64	57.1%	78.5%	89.4%
Structure 3	2	64	60.1%	81.5%	92.1%
Structure 4	1	64	40.8%	54.2%	58.6%
Structure 6	3	128	42.7%	50.6%	59.8%

Finally, we choose Structure 1 as the final model structure, and obtain corresponding translation results for 500 words in the standard special name database through phonetic symbol generation, syllable division, and conversion to Chinese characters. The accuracy obtained by comparison is approximately 92.7%.

The minimum entropy algorithm original corpus is from the phonetic symbols in the English dictionary, and 300 of them are selected for artificial syllable division as the test set. In the experiment, the average entropy of the original phonetic corpus decreased from 3.28 bits to 3.02 bits, and the probability distribution function of the final syllable set was obtained. The syllables obtained by experiments were compared with the syllables of 300 reserved special names that were manually divided. The accuracy rate is 93.3%.

3.3.2. Automatic Evaluation of Translation Results

This paper automatically evaluates the accuracy of translation results of all English geographical names. The main parameters (partial) are shown in Table 6.

Table 6. Evaluation parameters of translation results (partial).

Geographical Name	Translation Result	$acc_{translit}$	$SentTree_{Score}$	$acc_{template}$	Accuracy(Acc)
Sandsfoot Castle	桑德斯富特城堡	0.9	−14.43	0.96	0.93
North Walsham Railway Station	北沃尔沙姆火车站	0.9	−18.97	0.70	0.85
Solent Hotel and Spa	索伦特酒店及斯帕	0.9	−30.54	0.23	0.59

The lexical structure of geographical names of different lengths differs in complexity and will contribute differently to the evaluation indicators. To make the machine evaluation more accurate, the evaluation index values of the three groups of place name data, including two-word, three-word and multi-word names, were normalized, and the thresholds were set after analysis to determine the reliability interval. The distribution histogram of evaluation indicators for each group is shown in the Figures 10–12.

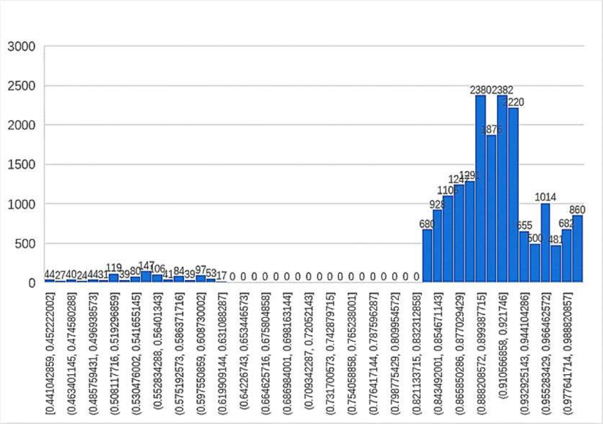

Figure 10. Distribution histogram of evaluation index valued of two-word geographical names.

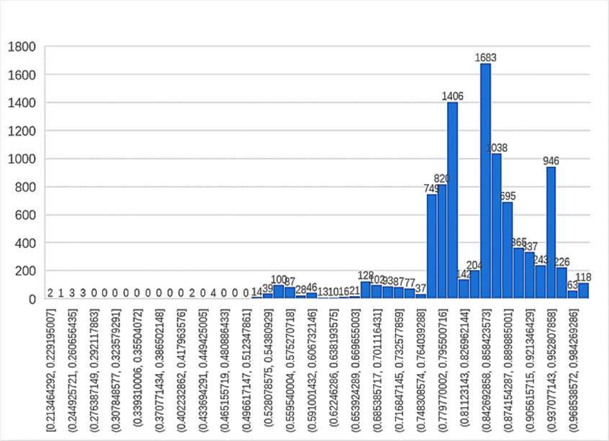

Figure 11. Distribution histogram of evaluation index values of three-word geographical names.

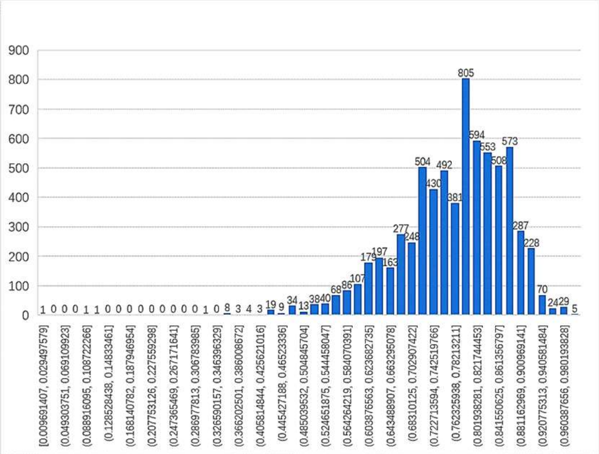

Figure 12. Distribution histogram of evaluation index values of multi-word geographical names.

According to the histogram of the two-word geographical name evaluation index, it was found that the index declined before 0.83; therefore, the threshold value of this group was set at 0.83, and the unreliable rate of the translation results was 1034/19334 = 5.34%. In the histogram of the three-word geographical name evaluation index, although there was a brief decline between 0.764 and 0.84, there was still a peak after 0.764; therefore, the threshold was set at 0.764, and the unreliability rate of the translation result was 912/9948 = 9.17%. In multi-word geographical names, the distribution in this group was similar to the right-deviation distribution of a skewed distribution. Because the skewed distribution is not suitable for statistical analysis, this paper converts it to a normal distribution by the Box–Cox variation (λ = 1.39), whose expected value and standard deviation are (μ,δ) = (0.623,0.112). Finally, the μ-2δ is selected as 0.399, and the threshold is 0.654 after reduction. The unreliability rate of the translation result is 909/6983 = 13.02%. The translation results of unreliable geographical names are shown in Table 7.

Table 7. Unreliable geographical name translation results (partial).

Geographical Name	Translation Result
Saint Just in Roseland	圣贾斯特在罗泽兰德中
North Lakes Hotel & Spa	北湖群酒店-斯帕
Holiday Inn Express Burnley m65 jct 10	伊克斯普雷斯-伯恩利-m65-杰西蒂-10-假日酒店
Premier Inn Manchester Airport m56 j6	普雷米厄旅馆曼彻斯特机场-m56 j6

4. Discussion

This paper designs a complete machine translation system for geographical names from the perspective of natural language processing and the theory of machine learning. There are three main innovations: (1) The system solves the problem of reasonable distinction between special names and general names in the translation of geographical names. (2) Using the theory of a cyclic neural network and minimum entropy, the phonetic symbol generation and phonetic division of English words are realized more efficiently and accurately. (3) A new idea is proposed for automatic quality inspection of translation of geographical names. The experimental results verify the reliability and scientific rigour of the method, which has important practical significance for the construction of global geographic information databases. In addition, some difficult problems based on rules are encountered in the experiment. The method needs to be optimized according to international translation standards or by listening to experts' opinions.

Author Contributions: Conceptualization, Xi Mao and Jizhou Wang; data curation, Hongkai Ren; formal analysis, Xi Mao; investigation, Hongkai Ren; methodology, Hongkai Ren and Xi Mao; project administration, Jizhou Wang and Weijun Ma; software, Hongkai Ren and Linyun Wang; supervision, Jizhou Wang; validation, Hongkai Ren and Weijun Ma; visualization, Hongkai Ren; writing—original draft, Hongkai Ren and Xi Mao; writing—review & editing, Hongkai Ren, Xi Mao, Jizhou Wang, Weijun Ma, and Linyun Wang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Projects of the Ministry of Natural Resources of China (12113600000018), the Key R&D Projects of the Ministry of Science and Technology of China (No. 2017YFB0503600), and the CASM (Chinese Academy of Surveying and Mapping) Research Fund (AR1912).

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

References

1. Roseredwood, R.; Alderman, D.; Azaryahu, M. Geographies of toponymic inscription: New directions in critical place-name studies. *Prog. Hum. Geogr.* **2010**, *34*, 453–470. [CrossRef]
2. Ivre, M.R.M.; Alencar, D. An ontological gazetteer and its application for place name disambiguation in text. *J. Braz. Comput. Soc.* **2011**, *17*, 267–279.
3. Aitchison, J.; Gilchrist, A.; Bawden, D. *Thesaurus Construction and Use: A Practical Manual*, 4th ed.; Aslib IMI: London, UK, 2005; pp. 133–167.
4. Menezes, Q.A. Dependency treelet translation: The convergence of statistical and example-based machine-translation. *Mach. Transl.* **2006**, *20*, 43–65.
5. General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China; National Standardization Administration Committee of China. *Guidelines for Chinese Translation of Foreign Language Geographical Names, GB/T 17693.1-2008*; Standards Press of China: Beijing, China, 2008; pp. 3–21.
6. Wang, J. Norms for general Chinese place name. *China Place Name* **2002**, *3*, 20–23.

7. Cheng, G.; Lu, X. Matching algorithm for Chinese place names by similarity in consideration of semantics of general names for places. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 404–410.
8. Wu, Y.; Schuster, M.; Chen, Z. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
9. Johnson, M.; Schuster, M.; Le, Q.V. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [CrossRef]
10. Shao, Y.; Nivre, J. Applying neural networks to English-Chinese named entity transliteration. In Proceedings of the Sixth Named Entity Workshop, Berlin, Germany, 12 August 2016; pp. 73–77.
11. Jiang, M.T.J.; Kuo, C.H.; Hsu, W.L. English-to-Chinese machine transliteration using accessor variety features of source graphemes. In Proceedings of the 3rd Named Entities Workshop (NEWS 2011), Chiang Mai, Thailand, 12 November 2011; pp. 86–90.
12. Knight, K.; Graehl, J. Machine transliteration. *Comput. Linguist.* **1998**, *24*, 599–612.
13. Brown, P.F.; Pietra, V.J.D.; Pietra, S.A.D. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.* **1993**, *19*, 263–311.
14. Chinese Character Reform Commission. *Chinese Pinyin Program*; People's Education Press: Beijing, China, 1956.
15. Xue, C.; Liu, J.; Li, X.; Dong, Q. Normalized-mutual-information-based mining method for cascading patterns. *ISPRS Int. J. Geo Inf.* **2016**, *5*, 174. [CrossRef]
16. Lee, S.; Lee, G.G. Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation. *Inf. Syst.* **2007**, *32*, 575–592. [CrossRef]
17. Wu, F.; Fu, K.; Wang, Y.; Xiao, Z. A graph-based min-# and error-optimal trajectory simplification algorithm and its extension towards online services. *ISPRS Int. J. Geo Inf.* **2017**, *6*, 19.
18. Xiong, D.; Liu, Q.; Lin, S. A dependency treelet string correspondence model for statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech, 23–23 June 2007; pp. 40–47.
19. Tanenhaus, M.K.; Boland, J.; Garnsey, S.M. Lexical structure in parsing long-distance dependencies. *J. Psycholinguist. Res.* **1989**, *18*, 37–50. [CrossRef] [PubMed]
20. Padó, S.; Galley, M.; Jurafsky, D. Machine translation evaluation with textual entailment features. In Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 30–31 March 2009; Volume 23, pp. 181–193.
21. Vardaro, J.; Schaeffer, M.; Hansen-Schirra, S. Translation quality and error recognition in professional neural machine translation post-editing. *Informatics* **2019**, *6*, 41. [CrossRef]
22. Specia, L.; Turchi, R.M. Special issue: Topics in machine translation evaluation || machine translation evaluation versus quality estimation. *Mach. Transl.* **2010**, *24*, 39–50. [CrossRef]
23. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
24. Pecina, P.; Schlesinger, P. Combining association measures for collocation extraction. In Proceedings of the ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17–21 July 2006.
25. Sorokina, A.; Aidana, K.; Zhenisbek, A. Low-rank approximation of matrices for pmi-based word embeddings. *arXiv* **2019**, arXiv:1909.09855.
26. Brent, M.R. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Mach. Learn.* **1999**, *34*, 71–105. [CrossRef]
27. GEOnet Names Server. Available online: http://www.geonames.org/ (accessed on 25 August 2019).
28. Yang, H. *Research on Template-Oriented Text Generation Technology for Machine Translation*; Northeastern University: Shenyang, China, 2011.
29. Quirk, C.; Menezes, A.; Cherry, C. Dependency treelet translation: Syntactically informed phrasal SMT. In Proceedings of the ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005.
30. Koehn, P. Statistical machine translation. *Desidoc. J. Libr. Inf. Technol.* **2009**, *30*, 25–32.
31. Watanabe, T.; Imamura, K.; Sumita, E. Statistical machine translation using hierarchical phrase alignment. *Syst. Comput. Jpn.* **2010**, *38*, 70–79. [CrossRef]

32. Sutskever, I.; Vinyals, O. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
33. Grundkiewicz, R.; Heafield, K. Neural machine translation techniques for named entity transliteration. In Proceedings of the Seventh Named Entities Workshop, Melbourne, Australia, 15–20 July 2018; pp. 89–94.
34. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
35. Redlich, A.N. Redundancy reduction as a strategy for unsupervised learning. *Neural Comput.* **1993**, *5*, 289–304. [CrossRef]
36. Darroch, J.N.; Ratcliff, D. Generalized iterative scaling for log-linear models. *Ann. Math. Stat.* **1972**, *43*, 1470–1480. [CrossRef]