Spatio- emporal Information Extraction in Literature Metadata

Qiuxia Du [1, 2, a], Hongguo Wang [1, 2, b], Zengzhen Shao [1, 2, a], Xin Fu [1, 2, a]

[1]Shandong Provincial Logistics Optimization and Predictive Engineering Technology Research Center, Jinan 250358, China

[2]School of Information Science and Engineering Shandong Normal University, Jinan 250358, China

[a]duqiuxia666@163.com, [b]729372827@qq.com

Abstract

In this paper, a hybrid HMM model of POS（Part of Speech）tagging,based on the complex representation of time information in literature metadata, is used to combine the regular expressions and time dictionaries to make the fuzzy time explicit in text, The time information in the literature metadata is annotated as a whole; Five rules of space-time pair recognition in Chinese text are proposed.The Spatiotemporal information Extraction Algorithm(SEA) is formed, which is based on the rule. Experiments on the time and location of bird species distribution in the study literatures show that this method can achieve good results in spatio - temporal information extraction of literature metadata.

Keywords

Metadata, Rule, Spatiotemporal Information Extraction, HMM.

1. Introduction

Time, space and attribute are the three basic characteristics inherent in things themselves. Time and space are important components to reflect the state of things and the process of evolution. Unstructured text contains a lot of time and space information, from the text to obtain unanalyzed, non-dominant space-time information is an urgent need for scientific research to solve the problem.

At present, in the aspect of text information extraction, the related researches focus on the independent element units such as names, place names, and organizational structure names, and the extraction of simple entity relations. For the extraction of temporal and spatial information has not attracted enough attention from scholars, some scholars from the perspective of natural language processing time information extraction and geographical name entity recognition conducted a preliminary exploration, Zhang Chunju et al [1] from the perspective of geospatial concept space-time semantic information , But the time information and the spatial information are separated and analyzed in the text, so there is no perfect method to extract the spatial and temporal information. In the aspect of time information extraction, the study of temporal information processing in English text is very mature. More typical for Allen put forward 13 kinds of time relationship theory, for the follow-up research laid a solid foundation [2]. Language description, the description of time information is flexible, semantic rich and so on. At present, the extraction of time information in Chinese text concentrates on the extraction of single time element. Time information is extracted and understood based on the analysis of time elements in Chinese and the form of time words, using the concept of time expression to map the time description to the time axis [3,4]. The extraction of information mainly focuses on the extraction of time information and the study of relative temporal relationships among different events, ignoring the determination of the time-space mapping relationship. In terms of the extraction of geographical names, the study focused on the identification of geographical names entities in the text. There are two ways to identify the entity names: 1) The method of combining gazetteer with manual rules has the advantages of easy implementation and high accuracy, but it can not solve the problem of new place name recognition and semantic diversity. 2) Machine learning method based on statistical model. This method is the main technique of current natural language processing. By analyzing the training corpus, we can set proper context eigenvector to train and identify the place names in the text. Hidden Markov Model, Support Vector Machine Model, Maximum Entropy Model, and Conditional Random Fields (CRFs) are also used for name recognition [4-7]. The development of the model is more and more complex, Cove model and the maximum entropy model, the single model is used in cascade, and the original basic model is modified, such as maximum interval hidden Markov model [5,9,10].

In this paper, the rules are combined with Hidden Markov Model (HMM) to mark the time, and then the time in the Chinese text is identified. In order to solve the problem of spatial-temporal information separation and separation, this paper proposes the matching rules of space-time pairs, and realizes the mapping of time-space relations, and extracts relatively accurate spatial and temporal information.

2. Time Information Identification Based on Hybrid Hidden Markov Model

2.1 Relevant models and algorithms

2.1.1 Hidden Markov Model

Hidden Markov Model (HMM) (see Fig. 1) is a double Markov stochastic process, which includes a Markov chain with state transition probability and a stochastic process of outputting observed values. Its state is uncertain or invisible, and only through the random sequence of observations can be demonstrated. An HMM consists of two layers: an observable layer and a hidden layer. The observable layer is the observed sequence to be identified, and the hidden layer is a Markov process, a finite state machine, where each state transition has a transition probability.

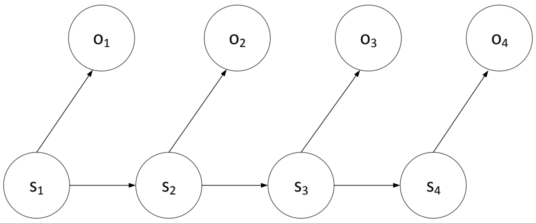

Fig. 1 Hidden Markov Model

An HMM can be seen as a five-tuple $\lambda=\{S, K, \Pi, A, B\}$, which $S=\{S_1, S_2, ..., S_N\}$ represents all the states in the model, N representing the number of states; $K=\{K_1, K_2, ..., K_M\}$ a set of state observations, M representing the number of observations; $\Pi=\{\pi_i = P(q_1 = S_i), 1 \le i \le N\}$ for the initial state of the set. π_i represents the probability of the state s_i as the initial state; $A=\left\{a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i), 1 \le i, j \le N\right\}$ is the transition probability matrix, and a_{ij} represents the probability that the state s_i transits to the state s_j ; $B=\left\{b_{i(k)} = P(O_t = K_k \mid q_t = S_i), 1 \le i \le N, 1 \le k \le M\right\}$ is the probability matrix of emission, and $b_{i(k)}$ is the probability of occurrence of s_i in state K_k .

2.1.2 Hybrid Hidden Markov Model（HHMM）

When the traditional HMM model is used to estimate the parameters, the probability of lexical emission of the current word w_i is only related to the current part of speech s_i . This calculation ignores the possible relationship between the current word and its adjacent postpositional part of speech. Backward dependency. Therefore, we propose an improved method for computing lexical probabilities. The emergence of words depends on its previous annotation s_{i-1} , its annotation si, and its later annotation s_{i+1} . In this paper, HHMM is composed of traditional HMM and Reverse Hidden Markov Model (RHMM) (see Fig. 2), which takes into account the effect of the latter state on the previous state.

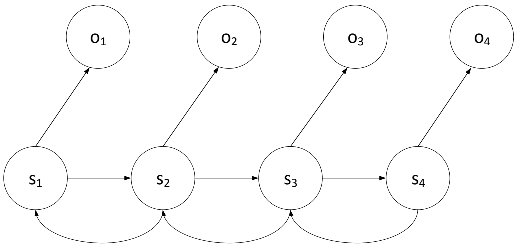

Fig. 2 Hybrid Hidden Markov Model

2.1.3 Maximum Likelihood（ML）

Given an observation sequence $o = o_1 o_2 ... o_T$, how to set the λ value to get the maximum $P(o|\lambda)$, the hidden Markov model parameter calculation problem usually use the maximum likelihood estimation [11], the algorithm is: If a test of n Possible outcomes $A_1,...,A_n$,, is to do a test, if event A_i occurs, then that event A_i in the n possible results in the probability of the largest. A simple sentence is used to describe the probability that an event (should) occurring in one trial has a greater probability. The following formula is used to calculate the parameter values using the ML algorithm:

The initial state probability calculation formula:

$$\pi_i = {Init(i)} \Big/ {\sum_{j=1}^{N} Init(j)} \qquad 1 \le i \le N$$

In all labeled training samples, $Init(i)$ is the number of sequences in which S_i is the initial state and $\sum_{j=1}^{N} Init(j)$ is the sum of the number of sequences in all states in the initial state.

State transition probability calculation formula:

$$a_{ij} = {C_{i,j}} \Big/ {\sum_{k=1}^{N} C_{i,k}} \qquad 1 \le i, j \le N$$

$C_{i,j}$ is the number of transitions from S_i to S_j, and $\sum_{k=1}^{N} C_{i,k}$ is the sum of the number of transitions from state S_i to all states.

The probability of emission calculation formula:

$$b_j(k) = {E_j(k)} \Big/ {\sum_{i=1}^{M} E_j(i)} \qquad 1 \le j \le N, 1 \le k \le M$$

$E_j(k)$ is the number of occurrences of O_k in S_j, and $\sum_{i=1}^{M} E_j(i)$ is the total number of occurrences of O_k in S_j.

2.1.4 R-Viterbi

For a particular HMM and a corresponding observation sequence, it is often desirable in practical applications to find the most likely hidden sequence state for generating this sequence. The problem to be solved in named entity recognition is how to find the most suitable tag sequence for a sentence, namely decoding. The commonly used decoding method is the Viterbi [12] algorithm, which belongs to the dynamic programming algorithm. The idea is to decompose the problem, solve the most basic sub-problems first, and then extrapolate the optimal solution of the better sub- The optimal solution of the whole problem is obtained, that is to get the best named entity marking sequence [13,14].

Definition 1.1: The Viterbi variable $\varphi_t(j)$ is the maximum probability that the HMM travels from a certain path to S_t at time t and outputs $o = o_1 o_2 ... o_T$.

The steps of the Viterbi algorithm are as follows:

Step 1: Calculate the initial local optimal function using the following formula:

$$\varphi_1(i) = \pi_i b_i(y_1) \qquad 1 \le i \le N$$

$$\theta_1(i) = 0$$

Step 2: Determine the recursion of recursive functions for solving local optima:

$$\varphi_t(j) = \max_{1\le i\le N}\left[\varphi_{t-1}(i)a_{ij}\right] b_j(y_t)\ 2 \le t \le T, 1 \le j \le N$$

$$\theta_1(j) = \arg\max_{1\le i\le N}\left[\varphi_{t-1}(i)a_{ij}\right]\ 2 \le t \le T, 1 \le j \le N$$

Step 3: Calculate the best state of the last observation:

$$P^* = \max_{1\le i\le N}\left[\varphi_T(i)\right]$$

$$q^*_T = \arg\max_{1\le i\le N}\left[\varphi_T(i)\right]$$

Step4: back before the observation should be the best state:

$$q^*_t = \theta_{t+1}(q^*_{t+1})\ t = T-1, T-2, \dots, 1$$

Here, the direction pointer $\theta_1(j)$ is used to record the path of the optimal sequence, and T is the number of observed values produced in the sequence of observations. According to the $\theta_1(j)$ back-off pointer from the current best state q^*_{t+1} can be obtained from the previous observation of the best state.

For the observation sequence $O = (O_1, O_2, \dots, O_T)$, when decoded with the Viterbi algorithm, $P(q_{i+1} \mid q_1, q_2, \dots, q_i) = P(q_{i+1} \mid q_i)$, that is, the state of O_2 is determined by the state of O_1. For the inverse hidden Markov model, the traditional decoding algorithm (Viterbi algorithm) is improved as follows: For a class of observation sequence $O' = (O_T, O_{T-1}, \dots, O_1)$, the state q_{t+1} at time $t+1$ is determined by state q_t at time t,That is, the state of O_1 is determined by the state of O_2. the decoding process of observation sequence $O' = (O_T, O_{T-1}, \dots, O_1)$ can be understood as follows: for the reverse process of observation sequence $O = (O_1, O_2, \dots, O_T)$, state q_t at time t is determined by state q_{t+1} at time $t+1$, i.e., $P(q_i \mid q_{i+1}, q_{i+2}, \dots, q_{i+n}) = P(q_i \mid q_{i+1})$, which reflects the backward dependency of the observed sequence .

2.2 HHMM-based sequence labeling method

2.2.1 Data preprocessing

In this paper, the NLPIR Chinese word segmentation system (ICTCLAS2014), which is selected by the Chinese word segmentation tool, is used to segment and label the initial text, and the word segmentation is selected according to the POS tagging. The word set after each text processing is used as the next step data source.

2.2.2 Model establishment

Firstly, the training corpus is preprocessed, including the removal of punctuation, and the corresponding word segmentation and POS tagging of the corpus. After processing, a document with POS tagging information is generated. To facilitate the calculation and removal of noise data, (N), the verb (v), and the adjective (a) are not subdivided for simplification; except for verbs, nouns, adjectives, etc., the following modifications are made: Other words (such as onomatopoeia, punctuation, etc.) are marked as o, according to the training process of mixed HMM model is as follows:

Step1: initialize the HMM, determine the number of state model, initialization model parameters. The state set is the label for each word, which is n (object name), ns (place name), t (time word), v (verb), a (adjective), r (pronoun), m (number) D (adverb), p (preposition), u (auxiliary word), o (other words);

Step2: Scan text in positive sequence, convert the tagged metadata into a sequence of packets according to the format and delimiters. Each packet is tagged with the previous step symbol.

Step3: Train the HMM model to calculate the HMM parameters $\lambda = (A, B, \pi)$ using the ML algorithm [11] in groups;

Step4: initialize RHMM, determine the number of state model, initialization model parameters, the state set and step1 the same.

Step5: reverse scan text, the same method and step2;

Step6: Training RHMM model to group as a unit, the application of ML algorithm to calculate RHMM parameters $\lambda=\left(A', B', \pi\right)$.

2.2.3 Improved Viterbi algorithm

In order to facilitate the calculation, this paper assumes that the words in the test text are only related to the sentences in which the words are located. First, the test text is punctuated by punctuation marks "," "." "!" "?", So that the test text becomes a single sentence; then these sentences are segmented to find the best labeling sequence. The algorithm is described as follows:

Step1: Training sample pre-processing. Scanning text, using comma, period, question mark, exclamation point (in brackets, the comma within the quotation marks excluded), more than three spaces and other delimiter information text segmentation into text block sequence;

Step2: Calculates the maximum probability path for positive sequence text. Combined with the training part of the output of the HMM model parameters $\lambda=\left(A, B, \pi\right)$, using Viterbi algorithm;

Step3: Look up the state of HMM on the maximum probability path and output the mark sequence.

Step4: Calculates the maximum probability path for reverse text. Combined with the RHMM model parameter $\lambda=\left(A', B', \pi\right)$bbb output by the training part, is calculated by R-Viterbi algorithm;

Step5: Look up the state of the RHMM on the maximum probability path, and output the markup sequence.

2.3 Time Information Identification

2.3.1 Time information classification

From the macroscopic points, time information can be divided into explicit time information and recessive time information. Such as "August 2016", "this summer" and so on, this type of time information can be easily identified, and hidden time information such as "when the swallows fly south", "tree germination" Etc., this kind of text information describes an event, but contains a wealth of time information, therefore, identify and identify the hidden time information is the key to extracting time information. At the same time, it is necessary to analyze the syntactic components in time information, including time phrases (TPP), temporal adverbs (15), time phrases (TNP) It is important to identify and mark this information for the extraction of time information.

Because time is a more complicated concept, there are stationary time points and continuous time periods, which bring certain complexity to extracting time information. In this paper, time information classification shown in Figure 3:

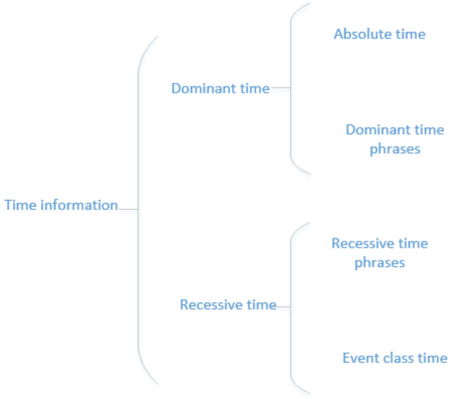

Fig. 3 Time information classification

2.3.2 Time information based on Regular expression and Time dictionary identification

Because ICTCLAS is used to extract time information, it can only extract single time element information, can not form a complete time expression, and can not support time reasoning and canonical expression. In this paper, we establish the time dictionary table for the fuzzy time in the text (see Table 1-1) to get the time information to the month.

Regular Expressions

The time involved in document metadata has three aspects: publication time, time in the title, and time in the summary.

Corresponding time recognition is also divided into three aspects, which published in the literature metadata in a specific representation. Such as: {Year}: 1986; less time in the title; the most and most important part is the time expression in the abstract, this paper uses HHMM model and regular expression matching method to identify. The specific regular expression is as follows:

[1]. Release Time Extraction Mode:

```
@" ( [{][Y][e][a][r][}][:][ ][0-9][0-9][0-9][0-9] )";
```

[2]. Time extraction mode in title and summary

a.Point Time Mode:

```
@" (\W+同年|同期/d)?(\w+\ /t) + ((\w+\ /w)?(\w+\ /p)?(\w+\ /t)?(\w+\/d)? (\w+\/f)? "
```

The part of speech of the time expression that this expression contains includes:

1.Time noun Such as “ 一九九七年/t 十二月/t 三十一日/t”

2. Time adverb + time noun | time noun + time adverb. For example,“上年/t 同期/d；同年/d 9月/t 10日/t”

3. Time boundary word + time noun + time boundary word. Such as“一九九七年/t 底/f；8日八晚上/t 到/p 9日/t”

b. Paragraph time mode:

Numeral + quantifier form, such as three days, two months and so on. Its regular expression:

```
@"((\w+/m)\S+)+(个/q)?\S+(年|月|{世纪|年代|天|星期|小时|周年|岁|秒|年度|周|载}+(/q|/n)+"
```

Time dictionary

The purpose of the time dictionary (see Table 1) is to associate the time word with the interval of the time axis and to find the appropriate value for the item in the time dictionary during the normalization process. We add the time words in the time dictionary, such as "the first month, Qingming, spring", etc., and assign the necessary attributes for each time word, the following is the time type and the corresponding relationship between its properties:

Specific time:

Ching Ming: date = xxxx / 04/05

Mid - Autumn: date = xxxx / oszl5L

Indicates the range of time:

Spring: date = xxxx / 03/01 ~ xxxx / 05/31

Table 1. Time dictionary table

Words of Time	Month	Words of Time	Month
正月、孟春、寅月、首春、元阳、初月、开岁、初春、小寒、大寒	January	孟秋、申月、巧月、首秋、初秋、兰月、瓜月、凉月、小暑、大暑	July
仲春、卯月、仲月、杏月、丽月、花朝、立春、雨水	February	仲月、酉月、中秋、正秋、桂月、爽月、立秋、处暑	August
季春、辰月、暮春、杪春、桃月、蚕月、惊蛰、春分	March	季秋、戌月、暮秋、菊序、霜序、菊月、咏月、白露、秋分	September

孟夏、巳月、清和、槐序、槐月、麦月、麦秋、清明、谷雨	April	孟冬、亥月、初冬、良月、阳月、开冬、寒露、霜降	October
仲夏、午月、蒲月、榴月、中夏、立夏、小满	May	仲冬、中冬、子月、畅月、复月、龙潜、雪月、冬月、大吕、立冬、小雪	November
季夏、未月、暑月、荷月、暮夏、芒种、夏至	June	季冬、丑月、严冬、嘉平、暮平、临月、腊月、风杪、残冬、冰月、岁暮、大雪、冬至	December
春季、初春、早春、阳春、芳春、暮春	March to May	秋季、初秋、中秋、暮秋、三秋、金秋、九秋、清秋、高秋、霜秋、霜天	September to November
夏季、初夏、中夏、暮夏、九夏、盛夏、炎夏、三夏	June to August	为冬季，冬季又包含：初冬、中冬、寒冬、九冬、暮冬、雪冬、冷冬、隆冬	December to next February

2.3.3 Time information Recognition algorithm

The time information recognition algorithm (TRA) based on regular expression and time dictionary is as follows:

Step1: Input non-empty sentence S;

Step2: If S does not contain time information, go to step1, otherwise go to step3;

Step3: if the time-phrase T conforms to the absolute time feature, go to step4; if T meets the fuzzy time feature, go to step5; if T does not fit either, and T is a mixture of the first two cases, And go to step3;

Step4: If the full format (xxxxxxxxx) is met, add the reference time stamp Tc and decompose Tc to (TY, TM, TD). Use regular expression to identify the time, and mark the time information as " T "; if it does not meet the full format, according to the maximum granularity of T, Tc inside the larger unit to complete the value; go to Step6;

Step5: T with the time dictionary table to match, get a relatively accurate time; go to step6;

Step6: the location of the time to combine the time, the overall time marked as "/ t", go to step2;

Step7: If the text ends, the algorithm terminates; otherwise go to step1;

After the time information is identified, the time information in the text is annotated as a whole, and the results are compared with those marked with ICTCLAS.

Table 2. The time information labels the results of the comparison

This article's time-tagging results	Results of ICTCLAS segmentation software
{/w Abstract/ws }/w :/w </w 正/a >/w 烟台市/n 鸟类/n 资源/n 普查/v 从/p 2014年5月/t 开始/v，/w 至/v 2016年6月/t 结束/v 。	{/w Abstract/ws }/w :/w </w 正/a >/w 烟台市/ns 鸟类/n 资源/n 普查/v 从/p 2014/m 年/nt 5/m 月/nt 开始/v ，/w 至/v 2016/m 年/nt 6/m 月/nt 结束/v 。
{/w Abstract/ws }/w :/w </w 正/a >/w 2016年4月/t，/w 我们/n 来到/v 信阳/ns 南湾/ns 林场/n 的/u 鸟岛/n 。/w 鸟岛/n 位于/v 距/v 信阳/ns 市/n 西南/nd 7/m 公里/q 处/v 的/u 南湾湖/ns 中/nd ，/w 面积/n 约/v 50/m 公顷/q 。	{/w Abstract/ws }/w :/w </w 正/a >/w 2016/m 年/nt清明/nt 时节/nt，/w 我们/n 来到/v 信阳/ns 南湾/ns 林场/n 的/u 鸟岛/n 。/w 鸟岛/n 位于/v 距/v 信阳/ns 市/n 西南/nd 7/m 公里/q 处/v 的/u 南湾湖/ns 中/nd ，/w 面积/n 约/v 50/m 公顷/q 。

As can be seen from the table above, this paper annotated method makes up for the shortcomings of recognizing single time element by using word segmentation software. Through improved labeling method, the complete time expression can be identified and the time information is labeled as A whole, the found time and place the relationship laid the foundation.

3. Rule - based temporal and spatial information extraction

The spatio-temporal information extraction process is shown in Fig.4.

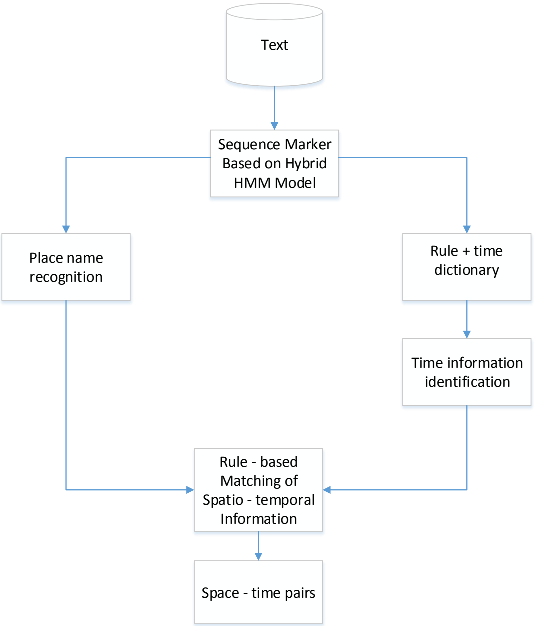
Text
Sequence Marker Based on Hybrid HMM Model
Place name recognition
Rule + time dictionary
Time information identification
Rule - based Matching of Spatio - temporal Information
Space - time pairs

Fig. 4 Temporal information extraction process

The spatial and temporal information concerned in this paper refers to extracting the time information and the location information from the Chinese text and matching them. The above research has identified the time information in the text, and I have given the method of site identification based on mixed HMM [16], and the next work is to identify the time and place (Extraction of objects such as Table 3), and to ensure that the time and space matching, matching refers to the text appears multiple times a number of locations or only the location of the time there is no time to find the correct time and place correspondence. The recognition of temporal and spatial relations needs to classify the temporal expressions and time-space pairs of the names and entities in the sentence, and then reproduce the relative exact time and place where the event occurs. Since the space-time pairs are not only for the two specific words, so in the identification of time and place relations feature extraction not only to consider the lexical features, but also to take into account the time and place of the relationship between features.

Table 3. Temporal and spatial information extraction object.

Text
{/w Abstract/ws }/w :/w </w 正/a >/w 烟台市/n 鸟类/n 资源/n 普查/v 从/p 2014年5月/t 开始/v，/w 至/v 2016年6月/t 结束/v 。
{/w Abstract/ws }/w :/w </w 正/a >/w 2016年4月/t，/w 我们/n 来到/v 信阳/ns 南湾/ns 林场/n 的/u 鸟岛/n 。/w 鸟岛/n 位于/v 距/v 信阳/ns 市/n 西南/nd 7/m 公里/q 处/v 的/u 南湾湖/ns 中/nd ，/w 面积/n 约/v 50/m 公顷/q 。

3.1 Feature extraction

For the purposes of the following description, the following definitions are made:

Definition 2.1: $s_{time\_site}\{wt_1,...,wt_m\}$ is a sentence containing time and place or contains only place, where $wt_i, 1 \le i \le m$ is the ith term in the sentence s_{time_site}, $wt_j = wt_{time}$ is a time expression in s_{time_site}, $wt_k = wt_{site}$is a place in s_{time_site}. (wt_{time}, wt_{site}) is called a space-time pair.

In temporal and spatial relation extraction, lexical features include word features, context features and verb features.

Definition 2.2: word characteristics: that is, space-time on all the words, that is, (wt_{time}, wt_{site}) itself.

Definition 2.3: Prepositional Phrase Features: All words contained in a prepositional phrase, such as "在2016年5月对烟台的10多种鸟类进行了研究", "在"and "对"are prepositional features that contain place name entities or time expressions. If the time or place is in the prepositional phrase ph_{prep}, the prepositional feature is expressed as $PRE=\{w \mid w \in ph_{prep}, w \neq wt_{time}, w \neq wt_{site}\}$.

Definition 2.4: Syntactic relational features: Syntactic relationships between temporal expressions and place name entities in sentences. This feature can be expressed as: whether in a sentence, whether in the same clause, whether in the same phrase. This feature is represented by the variable i_s.

Definition 2.5: Time and place distance characteristics: The number of words contained between the time expression and the location, denoted by the variable d.

Definition 2.6: Types of temporal expressions: The type of temporal expression in space-time alignment, denoted by the variable t. This feature includes point time and segment time

Algorithm: The time-site feature extraction algorithm (TSEA)

Input: The time-space pair (wt_{time}, wt_{site}) in the sentences $s_{time\_site}\{wt_1,...,wt_m\}$, s_{time_site} containing the time expressions and place names.

Output: The set of time - space pairs (wt_{time}, wt_{site}) is F_{time_site}

Algorithm steps:

Step1: Initialize the feature set: $F_{time\_site}=\phi$

Step2: extract (wt_{time}, wt_{site}) word feature and add feature set:

$$F_{time\_site}=(wt_{time}, wt_{site}) \cup F_{time\_site}$$

Step3: extract the prepositional phrase feature of (wt_{time}, wt_{site}) and add the feature set:

$$F_{time\_site}=PRE \cup F_{time\_site}$$

Step4: extract the (wt_{time}, wt_{site}) syntax characteristics and join the feature set:

$$F_{time\_site}=i\_s \cup F_{time\_site}$$

Step5: extract the time and location of (wt_{time}, wt_{site}) from the feature and add feature set:

$$F_{time\_site}=d \cup F_{time\_site}$$

Step6: extract the time expression type characteristics of (wt_{time}, wt_{site}) and join the feature set:

$$F_{time\_site}=t \cup F_{time\_site}$$

Step7: Outputs the feature set F_{time_site}.

3.2 Rulemaking

Based on the extracted feature set, the following rules are formulated:

Rule 1: Part of Speech Feature Rule

$$\begin{array}{l} wt_{time} \in s_{time\_site}\{wt_1,...,wt_m\} \\ wt_{site} \in s_{time\_site}\{wt_1,...,wt_m\} \\ IF\ \ (wt_P \in (wt_{time}...wt_{site})\,\&\&\, wt_V \in (wt_{site}.....)) \\ THEN\ \{\text{Output}(wt_{time}, wt_{site})\} \\ \ ELSE\ \{\text{Rule2}\} \end{array}$$

If the current sentence contains ns (t) and t (ie time), if ns contains t (preposition) and t contains v (verb) , It is determined that the time and place match is successful.

Rule 2: Syntactic relation rules

$$wt_{time} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$wt_{site} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$IF\ \ (i\_s = 1\&\&(wt_{time}, wt_{site}) \in \text{The same phrase})$$
$$THEN\ \{\text{Output}(wt_{time}, wt_{site})\}$$
$$ELSE\ IF\ \ (i\_s = 1\&\&(wt_{time}, wt_{site}) \in \text{The same sentence})$$
$$THEN\ \{\text{Output}(wt_{time}, wt_{site})\}$$

The rule first determines whether the time and place are in the same sentence. If it is considered that the time and place match succeeds, if not, then it is judged whether the time and place are in a sentence, and if it is in the same sentence,

Rule 3: The shortest distance rule

$$wt_{time1}\dots wt_{timen} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$wt_{site1}\dots wt_{siten} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$IF\ \ (d(wt_{timej}, wt_{sitej}) < d(wt_{timek}, wt_{sitek}))$$
$$(1 \le j \le n, 1 \le k \le n)$$
$$THEN\ \{\text{Output}(wt_{timej}, wt_{sitej})\}$$

The rule is that if more than one place at the same time in a single sentence, the statistical time and place between the number of words, the number of words less time to identify the location match.

Rule 4: Prepositional Phrase Rules

$$wt_{time} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$wt_{site} \in s_{time\_site}\{wt_1,\dots,wt_m\}$$
$$IF\ \ (PRE = 1)$$
$$THEN\ \{\text{Output}(wt_{time}, wt_{site})\}$$

The rule is that if the time and place in a preposition phrase in the time and place that match the success.

Rule 5: Temporal reasoning rules

$$IF\ \ (wt_{time} = \varnothing\&\&wt_{site} \in s_{time\_site}\{wt_1,\dots,wt_m\})$$
$$THEN\ \{wt_{time} = \text{Article published time``}\{Year\}\text{''} - 2; \text{Output}(wt_{time}, wt_{site})\}$$
$$ELSE\ IF\ \ (wt_{time} = \text{Fuzzy time}\&\&wt_{site} \in s_{time\_site}\{wt_1,\dots,wt_m\})$$
$$THEN\ \{wt_{time} = \text{Time dictionary table mapping}; \text{Output}(wt_{time}, wt_{site})\}$$

The rule is that if the sentence contains no time information, then the article published time minus two years to get the time and place in the sentence match. If the sentence contains only the location and fuzzy time information (such as the late spring and early summer), then use the time dictionary table mapping time and place in the sentence match.

3.3 Spatial and temporal information extraction method

Based on the above five rules, the following gives the spatio-temporal information extraction algorithm:

Algorithm: Spatiotemporal information extraction algorithm (SEA)

Step1: Enter the sentence $s_{time\_site}\{wt_1,\dots,wt_m\}$

(1) If the sentence contains a time and a place, then go to step2;

(2) if the sentence contains multiple time multiple locations, then go to step4;

(3) If the sentence contains only the location, then go to step5;

(4) if there is no time and place the implementation of Step1.

Step2: the implementation of rule 1, the output of space-time (wt_{time}, wt_{site}). If the rule 1 condition is not satisfied, go to step3;

Step3: the implementation of the rules 4, the output of time and space on the (wt_{time}, wt_{site}), otherwise the implementation of Rule 2;

Step4: the implementation of rule 3, the output of space-time (wt_{time}, wt_{site});

Step5: the implementation of rules 5, the output time and space on the (wt_{time}, wt_{site});

Step6: If the text ends, the algorithm terminates; otherwise go to step1;

4. Experiments and Results analysis

The data collected from the CNKI database include: Egret 2000, White Wagtail 1500, Bulbul 1400, Sparrow 2000, Blackbird 1300, and Beadstock. The experiment was developed in the Java language in the Eclipse development environment. In order to better evaluate the effectiveness of the algorithm, the data of the Chinese Bulbul and White Wagtail are used as the training data for other test data.

Experiment 1:

The results of HHMM-based method are compared with those of traditional HMM-based sequence tagging methods. The time recognition results (including the publication time) are listed in Table 4. The three common evaluation criteria are recall rate (R), accuracy rate (P) And F1 (Recall rate and accuracy of harmonic mean) to evaluate the performance of the algorithm, c on behalf of the total number of time, m represents the total time identified, n on behalf of the recognition of the right time, t on behalf of recognition error time. Then the following formula:

$$R = \frac{n}{c} \qquad P = \frac{n}{m} \qquad F_1 = \frac{2 \times P \times R}{P + R}$$

Table 4. Identification results (eg egrets)

Method	Total time （c）	Total number of times identified (m)	Identify the correct time (n)	Time to identify errors (t)
Rule+HMM	3650	3356	3080	276
Rule+HHMM	3650	3574	3341	233

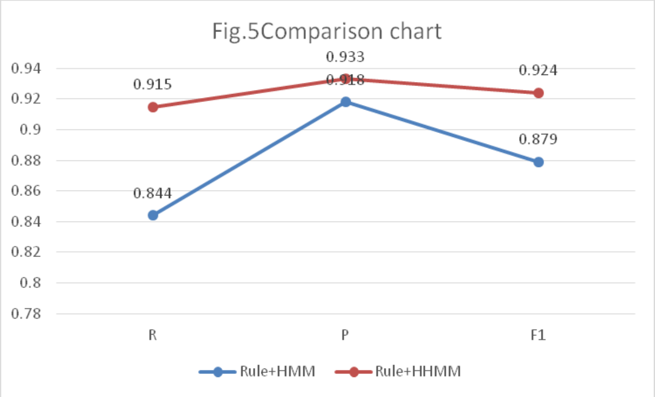

Figure 5. The results of HHMM-based method

As can be seen from Figure 5, the rule-based and HHMM-based methods have significantly higher recall and F1 values than rule-based and traditional HMMs, although there is an increase in accuracy but an increase in the accuracy of time information Not much. Therefore, the method of this paper is effective for time information recognition.

Experiment 2:

After the time information in the text is marked twice, combined with the former rules of spatial and temporal information extraction, the time-space results are as follows:

Table 5. Space-time pairs of extraction results

Bird 's name	Egrets	Sparrow	Blackbird	Bead neck doves
Extract the results	2012	2020	1302	1105

The extracted space-time is generally more than the number of texts. The reason, on the one hand, is that because there are many times in the text and multiple locations exist in different sentences, and multiple time and multiple locations can cross-match, that has its own space-time pairs. On the other hand there is a place in a sentence but there are multiple times and in the same sentence, in the matching process, each time with the location of the match, so there will be a number of space-time pairs. For example, the following summary section: “{Abstract}: 先后于2008年10月和12月、2010年3～4月和7～8月、2012年4～5月、2013年4～5月、2015年1月和2016年5月对广西北部湾涠洲岛的鸟类资源进行了调查,在该保护区内共记录到鸟类186种,隶属16目52科。” This sentence can be extracted out of eight space-time pairs.

Since there is no clear evaluation criterion for the extraction of spatio-temporal pairs, an Egret's text is taken as an example, and the 1000 texts of Egretta were manually extracted, and then compared with the results of this paper.

Table 6 Time-space pairs were extracted to compare the results

	Egret (1000 items)	Extraction right	Extraction error
Artificial extraction results	1008	1008	0
Rule-based approach	1020	960	60

From Table 6, it can be calculated that the accuracy of the method is 95.2%, and it has a good effect on the extraction of spatiotemporal information.

5. Conclusion

In this paper, a hybrid HMM is used to mark the double part of speech, and the regular expression is used to improve the F1 value of identifying the time information from the text. By using the five rules of time and place matching, we get the extraction method of time and space pairs, and extract the time and space information from the text to achieve high accuracy. Because there is no time for the text to extract only a simple use of article published time to match, to be in the future research gradually improved. The extraction of spatiotemporal information has a good application effect on the process of mining the implicit event in the text, and it is worth further study.

References

[1] Zhang Chunju.Chinese text in the event space and attribute information analysis method [D]. Nanjing Normal University, 2013.

[2] Allen J F. Towards a General Theory of Action and Time [J]. Artificial Intelligence, 1984, 23 (2): 123-154.

[3] Zhong Zhao-man, Li Cun-hua, Qiao Lei, et al.Efficient Web News Publishing Time Extraction Method [J] .Small Micro-Computer System, 2013, 34 (09): 2085-2089.

[4] Zhang Yuanpeng, Dong Jiancheng, Zhou Huiling, et al.Time information extraction in H7N9 event based on HMM [J]. Chinese Journal of Digital Medicine, 2015, 10 (10): 23-26.

[5] Zhang Yan. Hidden Markov Model based on the Chinese information extraction algorithm [D]. Liaoning University of Science and Technology, 2014.

[5] HE Yan-xiang, LUO Chu-wei, HU Bin-yao.Genetic Naming Entity Recognition Method Based on Combination of CRF and Rule [J]. Journal of Computer Applications and Software, 2015, 32 (1): 179-185.

[7] Qian Jing, Zhang Jie, Zhang Tao.Study on Chinese Names and Names Recognition Methods Based on Maximum Entropy [J]. Small-sized Microcomputer System, 2006 (09): 1761-1765.

[8] SUN Hong, CHEN Jun-jie.Research on the method of Chinese name recognition based on double CRF and rules [J]. Journal of Computer Applications and Software, 2014 (11): 175-177.

[9] Cheng Xueqi, Wu Dayong, Li Jingyuan, et al. Naming Entity Recognition Method and System in Microblogging Message: CN, CN 103268339 A [P]. 2013.

[10] Yu Xincong, Li Honglian, Lv Xueqiang.Application of Maximum Entropy and HMM in Chinese POS Tagging [J]. Wireless Internet Technology, 2014 (11).

[11] Seymore K, McCallum A, Rosenfeld R. Learning hidden Markov model structure for information extraction [C] // AAAI-99 Workshop on Machine Learning for Information Extraction. 1999: 37-42.

[12] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. IEEE Transactions on Information Theory, 1967, 13.

[13] LIU Fang, ZHAO Tie-jun, YU Hao, YANG Mu-yun, FANG Gao-lin.Chinese chunking analysis based on statistics [J] Chinese Journal of Information, 2000,06: 28-32 + 39.

[14] Bai Tiehu.Statistics-based automatic corpus of Chinese words tagging method research and implementation [D]. Tsinghua University, 1992.

[15] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web [C] // International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc. 2007: 68-74.

[16] DU Qiu-xia, WANG Hong-guo, SHAO Zeng-zhen, FU Xin, LIU Yan-min.Study on Document Metadata Names Extraction Based on Hybrid HMM [J] .Journal of Computer and Digital Engineering, 2017,01.