Tibetan Person Attributes Extraction Based on Support Vector Machine

Li-li GUO¹,² and Yuan SUN¹,²,*

¹School of Information Engineering, Minzu University of China
²Minority Languages Branch, National Language Resource and Monitoring Research Center 100081, Beijing, China
E-mail: guolili0305@163.com
* sunyuan_cn@163.com
*Corresponding author

Keywords: Tibetan, Person attributes, Kernel function, Support vector machine

Abstract. Person attributes extraction is an important branch in Natural Language Processing. At present, Chinese entity attributes extraction method is mainly based on rules, statistics and machine learning methods, and has achieved good results. But the study of Tibetan entity attributes extraction has a great space. The paper uses person attribute, case-auxiliary word, verbs and other related meaningful words in Tibetan corpus as the features, and utilizes support vector machine to train and predict the classifier. It provides support for search engine, information security, machine translation and other researches.

Introduction

With the rapid development of Internet, a large number of electronic text information resources appear in front of people, and more and more online information in the form of multiple languages is released. How to get the needed information quickly and accurately becomes a major problem. There is an urgent need for some automated tools to help people quickly find the information that you really need in the mass of information resources. Research on Tibetan person attributes extraction can drive the development of Tibetan economy, science and technology, culture and other fields. Person attributes extraction are extracted person related attributes from Tibetan corpus, such as birth date, gender, nationality and so on. It is the transformation of unstructured text into a structured text to meet people's needs.

In Tibetan text resources and literature classification, text statistics [1] and entropy value calculation [2], the construction of document resource construction and dictionary, word segmentation method [3], person recognition, as well as Tibetan speech recognition [4] other fields have a considerable part of the progress. Tibetan person attributes extraction is still in its infancy, there is still a lot of work to be done. Typical methods in English are based on the feature vector method [5-6] and based on kernel function method [7-8]. There are two methods for the specific application of research in the Chinese language [9-10]. Deng Bo et al. [11] introduced the lexical semantic matching technology. Zelenko [12] used kernel function. Zhang et al. [13] designed a kind of compound convolution tree kernel function. Liu Kebin et al. [14] realized the automatic extraction system of Chinese entity relation based on kernel function. Natural Language Processing [15] method used in Chinese can be used in Tibetan information processing. However, the actual process of using must consider problems that Tibetan and Chinese compared to have a larger gap in the study of natural language processing.

Description Method

The method of entity relation extraction based on machine learning usually transforms the discrimination of entity relation into classification problem. This is generally divided into two phases: training phase and testing phase. The Tibetan corpus is obtained by configured crawler system from a Tibetan website, then selected articles about introducing the person introduction and
processed these sentences such as artificial segmentation, marking word class. Finally, the labeled corpus to vector is used as the input data to study and construct the classifier. In the testing stage, the trained classifier is used to classify the test data, and processing is shown in Figure 1. In the case of corpus preprocessing reaching a certain standard, the classifier selects the SVM with better comprehensive performance.

![Figure 1. Extraction flow chart.](image)

**Tibetan Corpus Processing**

The Tibetan corpus is obtained by configured crawler system from a Tibetan website, such as Wikipedia (Tibetan Edition), Kangba media network, China Tibetan middle school network, etc, then selected articles about introducing the person introduction and processed these sentences such as artificial segmentation, marking word class, as shown in example 1-2.

**Example 1:** 
\[<e1>ཇོ་ནང་ʚ་ར་ཐ།/nh</e1>་ནི<e2>/vཕོ/n</e2>རབས་ཤིག་ཡིན་།/w\]

**Example 2:** 
\[<e1>གཙང་ɧོན་ཧེ་ɻ་ཀ་ɻས་པའི་Ȅན་ཅན།/nh</e1>ནི<e2>/vǧང་གོ/n</e2>ཡིན་།/w\]

Vectoring of characteristics mainly uses keyword characteristics, annotation combination characteristics and marked characteristics of words nearby. High frequency keywords appear in a certain attribute category, with a high degree of recognition. There are many different Tibetan and Chinese and English. Tibetan is a post-predicate sentence. Verbs are the core of a sentence in Tibetan [16]. The case-marking of verb nearby is rich semantic role information. To a certain extent, the case-marking reflects the relationship [3] between the predicate and the subject in the sentence, and the appearance of these markers has certain rules. The case-marking is an important effect for Tibetan attributes extraction. So the case-marking and named entity marking in combination may be helpful to the classification effect. The marking characteristics of keywords nearby have a certain effect on improving the classification performance like the role of entity words.

**Attributes Extraction Based on SVM**

Because the data in the feature space can be divided into linear and some nonlinear, the support vector machine can be divided into linear SVM and nonlinear SVM. In 1963 SVM [17] is a new and very promising classification technology AT&TBell laboratory research team led by Vapnik. SVM is a pattern recognition method based on statistical learning theory, which is mainly used in the field of pattern recognition. Now SVM has been successfully applied in many fields (image processing, text and handwriting recognition, etc).

In the case of linear separable, SVM constructs a hyper plane \( H \) as in Eq. (1).

\[
w \cdot x + b = 0 \tag{1}\]

Formula: \( w \) is the weight vector, \( x \) is a feature vector, and \( b \) is a parameter. The hyper plane is separated from the positive and negative samples in the form of the maximum boundary. The hyper plane is constructed by finding the weight vector \( w \) and the parameter \( b \) to satisfy the conditions, as in Eq. (2) and (3).

\[
w \cdot x_i + b \geq 0, \text{ (Positive samples, } y = +1) \tag{2}\]

\[
w \cdot x_i + b \leq 0, \text{ (Negative samples, } y = -1) \tag{3}\]

when \( "w" \) reached the minimum. The \( x_i \) of the formula represents the feature vector of number \( i \) training sample; \( \|w\| \) represents weight vector \( w \) of the Euclidean norm; \( y \) is a sample class label.
After finding out the $w$ and the $b$, the class of the corresponding test samples is judged by the decision function, as in Eq. (4).

$$y_i = \text{sgn}(w \cdot x_i + b)$$ (4)

If the decision function value is +1, the sample is positive; otherwise, it belongs to the negative sample, as shown in Figure 2. It is introduced from the concept of the optimal classification surface of two kinds of linear separable cases. When it is extended to the two kinds of nonlinear situations, the optimal classification surface can be expressed as in Eq. (5) and (6).

$$\min_{w,b} \left( \frac{1}{2}w^Tw + C\sum_{i=1}^{n} \zeta_i \right)$$ (5)

$$y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, 2, \ldots, n$$ (6)

Among them, $x_i$ is the d-dimensional weight vector. $w$ is the d-dimensional weight vector, $(x_i)$ is the feature vector, mapped to high dimensional (possibly infinite dimensional) space function, $n$ is the number of samples. $n$ is the number of samples. $c$ is a penalty factor. By adjusting the value of the penalty factor, the best results are obtained. The discriminant function as in Eq. (7) can be obtained by solving the dual problem of the problem.

$$f(x) = \text{sgn}\left( \sum_{i=1}^{n} \alpha_i^* y_i K(x_i, x) + b^* \right)$$ (7)

Among them, $K(x_i, x) = \phi(x_i)^T \phi(x)$ is called the kernel function[21]. The common kernel functions are polynomial kernel function, radial basis kernel function and Sigmoid kernel function. The key of SVM is the kernel function. In this paper, the radial basis function as in Eq. (8) (RBF) is chosen as the kernel function, which set the parameter $\gamma = 1/k$ $k$ is the number of person attributes categories.

$$K(x_i, x) = \exp \left( -\gamma |x - x|^2 \right) \gamma > 0$$ (8)

This paper uses hierarchical SVM Hierarchical classification method first divides all categories into two sub categories, and then divides the sub class into two sub categories, which has been circulating until a single category is obtained. Hierarchical classifier only needs to construct $k - 1$ classifier. The structure is similar to binary tree, as shown in Figure 2.

![Figure 2. Hierarchical classification structure.](image)

**Experimental Results Analysis**

The corpus of each attribute is divided into 3 parts, with 6,144 sentences as the training data, the 3,080 sentence as the test data. During the experiment, the best experimental results are obtained by adjusting the feature vector, the kernel function and penalty factor of SVM under the premise of in the annotation and marking of Tibetan corpus to achieve the highest accuracy rate. The results of the experiment are shown in Table 1. The $F$ value of the father, mother, gender and other attributes is relatively high, and the nationality and occupation and other attributes of the $F$ value is relatively low. The key words such as father, mother and gender are more obvious and have high recognition,
and the characteristics of nationality and occupation have some difficulty in the stage of feature vector. So the relative correct rate is lower.

Table 1. Tibetan attributes extraction results.

<table>
<thead>
<tr>
<th>Attribute class</th>
<th>Precision(%)</th>
<th>Recall(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father</td>
<td>80.47</td>
<td>79.22</td>
<td>79.84</td>
</tr>
<tr>
<td>Mother</td>
<td>82.51</td>
<td>78.44</td>
<td>80.42</td>
</tr>
<tr>
<td>Date of birth</td>
<td>79.28</td>
<td>74.55</td>
<td>76.84</td>
</tr>
<tr>
<td>Place of birth</td>
<td>79.84</td>
<td>71.18</td>
<td>75.26</td>
</tr>
<tr>
<td>Gender</td>
<td>82.77</td>
<td>81.04</td>
<td>81.90</td>
</tr>
<tr>
<td>Death day</td>
<td>77.01</td>
<td>74.81</td>
<td>75.89</td>
</tr>
<tr>
<td>Nationality</td>
<td>70.79</td>
<td>69.87</td>
<td>70.33</td>
</tr>
<tr>
<td>Occupation</td>
<td>71.11</td>
<td>66.49</td>
<td>68.72</td>
</tr>
</tbody>
</table>

Hierarchical SVM to extract Tibetan person attributes as shown in Table 2. It provides support for search engine, information security, machine translation and other researches.

Table 2. Results of Song zanganbu's attributes extraction.

<table>
<thead>
<tr>
<th>Attribute class</th>
<th>Extract attribute value</th>
<th>Sentences that contain attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father</td>
<td>甸 སྐྱེ་གནམ་རི་ཨོང་བཙན་</td>
<td>甸 སྐྱེ་གནམ་རི་ཨོང་བཙན་བོའི་ཡབ་ནི་甸 སྐྱེ་གནམ་རི་ཨོང་བཙན་ཡིན།</td>
</tr>
<tr>
<td>Mother</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Date of birth</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Place of birth</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Gender</td>
<td>དབང་ བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Death day</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Nationality</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
<tr>
<td>Occupation</td>
<td>བོད་ཀྱི་ཕོ་རབས་ཤིག</td>
<td>甸 སྐྱེ་གནམ་པོ་ནི་བོད་ཀྱི་ཕོ་རབས་ཤིག་ཡིན།</td>
</tr>
</tbody>
</table>

Conclusions

This paper introduces a kind of using SVM method for Tibetan person attributes extraction. The methods of different feature selection are different effects for attributes extraction. At present, the corpus of person attributes extraction is not rich. The Tibetan data is relatively simple when compared with the test data of the Chinese in the experiment. And the number of labels, the expansion of the corpus content and inspection work is still further improved. The experimental results in a certain extent are obtained. In the study of Tibetan person attributes extraction is still great room for improvement.

Acknowledgments

This work is supported by National Nature Science Foundation (No. 61501529, No. 61331013), National Language Committee Project (No. YB125-139, ZDI125-36), and Minzu University of China Scientific Research Project (No. 2015MDQN11, No. 2015MDTD14C).

References


