International Journal of u- and e- Service, Science and Technology
Vol.9, No. 12 (2016), pp.81-86
http://dx.doi.org/10.14257/ijunesst.2016.9.12.07

Recognition of Person Name in Uyghur Text Corpus using Naïve Bayes

Abdurahim Mahmoud[1], Tashpolat Nizamidin[1], Peride Tursun[2] and Askar Hamdulla[2*]

[1]*Institute of Information Science and Engineering, Xinjiang University, China*
[2]*School of Graduate, Xinjiang University, Urumqi，China 830046*
askarhamdulla@sina.com

Abstract

This paper presents a novel approach to recognize person name in Uyghur corpus. The Recognition of a person name for Uyghur using Naive Bayes Classifier is a challenging task in intelligent computing. Uyghur person name recognition (UPNR) aims at classifying each word in a document into predefined target label (person name or others) in a linear and non-linear fashion. Some language specific rules are added to recognize person names. Moreover, some gazetteers and context patterns are added to increase its performance as it is observed that identification of rules and context patterns requires language-based knowledge to make the work better. We have used required lexical databases to prepare rules and identify the context patterns for Uyghur. Experimental results show that our approach achieves higher accuracy than previous approaches.

Keywords: *Uyghur, Person name, Naïve Bayes classification, Rules*

1. Introduction

The term "Named Entity", now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996) [1]. At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization, location names and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition" [2].

Person, organization and location names have different characteristics in Uyghur language; therefore, various entities were studied respectively. This paper mainly introduces the study of Person Name Recognition in Uyghur. From 2011 to 2015. While early systems were making use of handcrafted rule-based algorithms, modern systems most often resort to machine learning techniques. One of the first research papers in the field was presented by Li Jiazheng (2011) at The Journal of Chinese Information [3]. Li's paper describes a method for recognizing and translating Chinese names in Uyghur. It relies on both Uyghur and Chinese language models, in addition to using the traditional rule-based approach. The first statistical approach for Uyghur Person Name Recognition was presented by Askar Rozi (2013) at The Journal of Tsinghua University (Science and Technology) [4]. In his paper, conditional random fields (CRFs) was applied for recognizing Uyghur person names The agglutinative characteristics of the Uyghur language were used to determine the word, part-of-speech, word stem, suffix, first and last syllables, and the nearest verb as features. A greedy algorithm was used to select the best

* Corresponding Author

ISSN: 2005-4246 IJUNESST

feature templates for the recognition model. Jarulla Muhammad *et. al.*, (2014) presented a hybrid approach for automatic identifying of Uyghur person names at The Journal of Xinjiang University (Natural Science Edition) [5]. It realized identifying of candidate person names and eliminated ambiguity using statistical boundary model through analyzing the characteristics of Uygur person names, extracting feature sets and summarizing corresponding recognition rules. Abdurehim Mahmoud, Hussein Yusuf *et. al.*, (2013) presented a rule-based approach at the National Conference on Man-Machine Speech Communication which uses Dice Coefficient and Levenshtein Distance[6]. A letter-based fuzzy matching method was used for Uyghur person names, while the syllable-character conversion method which is inspired by the idea of machine translation was used for Chinese person names.

2. The Uyghur Text Corpus

For the real distribution of various person names in the Uyghur corpus, we count 197 news from <people.com.cn> Uyghur version. Including 111 news contains person names, so 86 news does not contain person names. the whole corpus consists of 14339 words, containing 231 person names, including 92 Uygur names (39.8%), 88 Han people names (38.1%), 23 Europe and American person names (10.1%), 12 Arabic person names (5.2%), 9 Russian person names (3.9%), and 7 other names (3%). From the distribution of all kinds of names can be seen, Uyghur names and Chinese names accounted for 77.9% in Uyghur news corpus, the rest of type of names accounted for 22.1%.

3. Naïve Bayes Classifier

The Naïve Bayes classifier is a Generative Model of supervised learning algorithms. It is a simple probabilistic classifier which is based on Bayes' theorem with strong and naïve independence assumptions between every pair of features. It is one of the most basic classifier used for text classification. Moreover, the training time with Naïve Bayes is significantly smaller as opposed to alternative methods such as Support Vector Machine (SVM) and Maximum Entropy (Max-Ent) classifiers. Naïve Bayes classifier is superior in terms of CPU and memory consumption as shown by Huang, J. (2003) [7]. Its performance is very close to SVM and Max-Ent classifiers.

The Multinomial Naïve Bayes classifier is suitable for classification with discrete features. The multinomial distribution normally requires integer feature counts; however, fractional counts such as Term Frequency and Inverse Document Frequency (tf-idf) will also work. Multinomial Naïve Bayes classifier is based on the Naïve Bayes algorithm. In order to find the probability for a label, this algorithm uses the Bayes rule to express P (label | features) in terms of P (label) and P (features | label). The Naïve Bayes classifier requires training data samples in the format: (xi, yi) where, xi includes the contextual information of the word/document (the sparse array) and yi, its class. Graphical representation of Naïve Bayes decoder is shoccccwn in Figure1. Here fi is ith feature of vocabulary (vi = xi) and P (f |yj) = P (xi = vi | yj) is the maximum probability that the input xi belongs to the class yj.

$$p(y_j \mid x_1, x_2, ..., x_n) = \frac{p(y_j)p(x_1, x_2, ..., x_n \mid y_j)}{p(x_1, x_2, ..., x_n)} \quad (1)$$

4. Training Data

The experimental data is consisted of labeled corpus of 11257 sentences. Among them, the training corpus has 10805 (186885 words) sentence, the test corpus has 1650 (21183 words) sentence which in the training corpus has 10359 person names, the test corpus has 2359 person name. All corpuses contain 12718 person names.

4.1. Features

It is mentioned the following set of features that have been applied to the UPNR task.

- Surrounding word is set to be +1 otherwise it is set to −1. This binary value used to all surrounding word feature.
- Person prefix word, if the prefix belongs to ‘abdu’, ‘gvl’, ‘memet’ *etc.*, then set to +1.
- Person suffix word, if the suffix belongs to ‘jan’, ‘eli’ *etc.*, then set to +1.
- IV (In vocabulary) Person name, if the word is in Uyghur person name dictionary. then set to +1.
- After stem segmentation, if the stem word is in Uyghur person name dictionary. then set to +1.

All positive words used in the training set are considered as +1 and rest of the words are considered as −1.

It is identified that various features may be considered to find out Person name in Uyghur language as mentioned below. Following the features many person names are identified. Also some rules are mentioned in this paper that is used for such purpose, as summarized below:

- An Uyghur word which is associated with its prefix or suffix word and its surrounding words *i.e.,* reis “chairman”, zungtung “president”, dedi “said”, ependi “Mr”, hanim “Mrs”, hojayin “boss”, dohtur “doctor” are used to identify person name.
- An Uyghur syllable which is like jan, han, gvl, ay, met, abdu, *etc.*, are used to identify person names. Some of the previous suffix are also used in Uyghur to identify person names, *e.g.*, -chi, -si.
- An Uyhgur word which is single syllaybe word like xi, lyu, huang, yao, li, *etc*., are used to identify Chinese person name in Uyghur language.

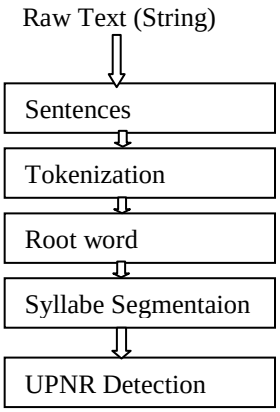

Figure 1. Flowchart of Finding UPNR

4.2. Suffix and Prefix

Some suffix and prefix alphabets are used to identify UPNR, which are mentioned in the features. Firstly a fixed length word suffix of the current and surrounding words are used as features.

4.3. Stem of Word

Morphological analyzer is used to find the stem of words by stripping suffix-prefix from a word.

4.4. The Algorithm

The proposed algorithm is used for finding the UPNR in the Uyghur corpus data. First, the entire Uyghur text corpus is entered by user in our proposed system, and then the process of UPNR is divided into seven steps which are described in the following algorithm.

Step 1: Enter a text

Step 2: Convert entire text into token by tokenization.

Step 3: Find stem word using morphological analysis.

Step 4: Compare each word with our valid features.

Step 5: Extract the features from each and every word.

Step 6: Compare each word with the training data set.

Step 7: Find the exact Uyghur person name.

5. Experimental Results and Evaluations

In this paper, the experimental evaluation criteria for accuracy (P), the recall rate (R) and the F1 value of 3 indicators, the formula (2), (3), (4) shown.

$$P = \frac{Actual\_taged\_UPN}{Taged\_as\_UPN} \times 100\% \quad (2)$$

$$R = \frac{Actual\_taged\_UPN}{UPN\_in\_corpus} \times 100\% \quad (3)$$

$$F1 = \frac{2 \times (P \times R)}{P + R} \times 100\% \quad (4)$$

This paper gives a 2 single feature test, to determine the context word window length. For the extraction of context word features, the window selection is very important, we determine the window length is 3 (previous word, the word and last word), as shown in Table 1.

Table 1. Context Window Experiment

Window Length	P/%	R/%	F1/%
$[W]_{-1}+[W]_0+[W]_1$	93.70	55.85	69.98
$[W]_{-2}+[W]_{-1}+[W]_0+[W]_1+[W]_2$	94.28	53.42	68.21
$[W]_{-3}+[W]_{-2}+[W]_{-1}+[W]_0+[W]_1+[W]_2+[W]_3$	94.23	50.19	65.49

In this paper, the characteristics of the context words as the basic features, , made 5 comparative experiments respectively, as shown in Table 2. Table 3, for the results of 7 comparative experiment.

Table 2. Comparative Experiment

Experiment	Feature template
1	$[W]_{-1}$、$[W]_0$、$[W]_1$
2	$[W]_{-1}$、$[W]_0$、$[W]_1$、$[Dic]_0$
3	$[W]_{-1}$、$[W]_0$、$[W]_1$、$[Stem]_0$、$[Suffix]_0$、$[Stem]_{-1}$、$[Suffix]_{-1}$、$[Stem]_1$、$[Suffix]_1$
4	$[W]_{-1}$、$[W]_0$、$[W]_1$、$[S_1]_0$、$[S_N]_0$、$[S_1S_2]_0$、$[S_{N-1}S_N]_0$、$[S_1S_2S_3]_0$、$[S_{N-2}S_{N-1}S_N]_0$

5	$[W]_{-1}$、$[W]_0$、$[W]_1$、$[C_1C_2]_0$、$[C_{M-1}C_M]_0$、$[C_1C_2C_3]_0$、$[C_{M-2}C_{M-1}C_M]_0$、$[C_1C_2C_3C_4]_0$、$[C_{M-3}C_{M-2}C_{M-1}C_M]_0$、$[C_1C_2C_3C_4C_5]_0$、$[C_{M-4}C_{M-3}C_{M-2}C_{M-1}C_M]_0$
6	$[W]_{-1}$、$[W]_0$、$[W]_1$、$[Stem]_0$、$[Suffix]_0$、$[Stem]_{-1}$、$[Suffix]_{-1}$、$[Stem]_1$、$[Suffix]_1$、$[S_1]_{-1}$、$[S_N]_{-1}$、$[S_1S_2]_{-1}$、$[S_{N-1}S_N]_{-1}$、$[S_1S_2S_3]_{-1}$、$[S_{N-2}S_{N-1}S_N]_{-1}$、$[S_1]_0$、$[S_N]_0$、$[S_1S_2]_0$、$[S_{N-1}S_N]_0$、$[S_1S_2S_3]_0$、$[S_{N-2}S_{N-1}S_N]_0$、$[S_1]_1$、$[S_N]_1$、$[S_1S_2]_1$、$[S_{N-1}S_N]_1$、$[S_1S_2S_3]_1$、$[S_{N-2}S_{N-1}S_N]_1$、$[C_1C_2]_{-1}$、$[C_{M-1}C_M]_{-1}$、$[C_1C_2C_3]_{-1}$、$[C_{M-2}C_{M-1}C_M]_{-1}$、$[C_1C_2C_3C_4]_{-1}$、$[C_{M-3}C_{M-2}C_{M-1}C_M]_{-1}$、$[C_1C_2C_3C_4C_5]_{-1}$、$[C_{M-4}C_{M-3}C_{M-2}C_{M-1}C_M]_{-1}$、$[C_1C_2]_0$、$[C_{M-1}C_M]_0$、$[C_1C_2C_3]_0$、$[C_{M-2}C_{M-1}C_M]_0$、$[C_1C_2C_3C_4]_0$、$[C_{M-3}C_{M-2}C_{M-1}C_M]_0$、$[C_1C_2C_3C_4C_5]_0$、$[C_{M-4}C_{M-3}C_{M-2}C_{M-1}C_M]_0$、$[C_1C_2]_1$、$[C_{N-1}C_N]_1$、$[C_1C_2]_1$、$[C_{M-1}C_M]_1$、$[C_1C_2C_3]_1$、$[C_{M-2}C_{M-1}C_M]_1$、$[C_1C_2C_3C_4]_1$、$[C_{M-3}C_{M-2}C_{M-1}C_M]_1$、$[C_1C_2C_3C_4C_5]_1$、$[C_{M-4}C_{M-3}C_{M-2}C_{M-1}C_M]_1$

Table 3. Experiment of Best Template Selection

Experiment	P/%	R/%	F1/%
1	93.70	55.85	69.98
2	92.30	62.22	74.33
3	94.02	66.53	77.92
4	93.70	71.57	81.16
5	92.56	73.46	81.91
6	85.86	94.90	90.15

In contrast with previous work, we have Askar Rozi et al proposed approch to recognition Uyghur names based on conditional random fields [1] as the baseline system. We made comparison experiment with baseline system. Results are shown in Table 4.

Table 4. Comparative Experiment

Experiment	P/%	R/%	F1/%
Baseline	90.03	82.96	86.35
Our method	85.86	94.90	90.15

6. Conclusions

Our work concentrated on recognizing person names from open-domain text, first we have showed that previous work of Uyghur person name recognition, because of complicated characteristics of Uyghur and lack of resources, then we introduced a method to construct Uyghur corpus, experiments based on Naive Bayes model have shown that the efficiency of our approach. In our model feature selection is done through manual assumption but implementation of technique in feature selection could be the future road map and then the experiment data is achieved by us from network, not open data, at the same time, scales of data have to be enlarged. And the features in the model can be tried and implemented in other agglutinative languages.

Acknowledgments

This work has been supported by Innovation Program for Excellent Ph.D. Candidates of Xinjiang University (XJUBSCX-2012010), the National Natural Science Foundation of China under grant of (61562081), and High Technology Research and Development Project of Xinjiang (201312103).

References

[1] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification”, Lingvistica Investigations, vol. 30, no. 1, **(2007)**, pp. 3-26.
[2] M. C. Andrew and W. Li, “Early Results for Named Entity Recognition with Conditional Random

Fields, Features Induction and Web-Enhanced Lexicons", In Proc. Conference on Computational Natural Language Learning, **(2003)**.
[3] J. LI , K. LIU, A. Mairehaba, Y. LV, Q. LIU and T. Yibulayin, "Recognition and Translation for Chinese Names in Uighur Language", JOURNAL OF CHINESE INFORMATION PROCESSING, no. 4, **(2011)**, pp. 82-87.
[4] A. Rozi, C. ZONG, G. Mamateli and A. Hamdulla, "Approch to recognition Uyghur names based on conditional random fields", Journal of Tsinghua University（Sci & Tech, vol. 53, no. 6, **(2013)**, pp. 873-877.
[5] J. Muhammad, T. Ibrahim and H. Omar, "Research of Uyghur Person Names Recognition Based on Statistics and Rules", Journal of Xinjiang University (Natural Science Edition), no. 3, **(2014)**, pp. 319-324.
[6] H. Yusuf, "The WEB Text Processing Algorithms for Public Opinion Analysis". Master Thesis of Xinjiang University of China, **(2014)**.
[7] J. Huang, J. Lu and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy", Data Mining, **(2003)**.

Authors

Abdurahim Mahmoud, Abdurahim Mahmoud Received his BSc degree in Electronics and Information System from Xinjiang University, Xinjiang, China in 1996, and MSc degree in Mechanical Design Theory from Xinjiang University, Xinjiang, China in 2007. He joined Xinjiang University as an assistant teacher in 1996. Currently, he is a doctoral student of computer science, his research direction is natural language processing.

Tashpolat Nizamidin, Tashpolat Nizamidin has received his B.E. degree in Electronics from Xinjiang University, China, in 2013. Currently, he is a M.S Student in Signal & Information processing in Xinjiang University. His research interest is Natural language Processing.

Palidan Tuerxun, Palidan Tuerxun received her M. S. degree in 1996 from Liaoning University, China and her Ph.D. degree in 2015 from Northwestern University, China. Since 1992, she has been working as a teacher at Xinjiang University, and since 2004, she was an associate professor in school of software of Xinjiang University. Her research interests are machine learning and Uyghur natural language processing.

Askar Hamdulla, Askar Hamdulla received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 160 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.

