Journal of Chinese Language and Computing 15 (4): (219-226)

Rule-based Person Name Recognition for Xinjiang Minority Languages

Gulila Altenbek
Department of Electrical and Information Science,
College of Information Science and Engineering,
Xinjiang University, Urumqi, Xinjiang 830046, P.R. China
E-mail: gla@xju.edu.cn

Abstract

Xinjiang multi-nationality name entity recognition is an important part in multi-language processing. In this paper, we analyze the patterns of Uighur and Kazak person names, and perform the name identity recognition using rule-based approach. We also propose and implement the rules for Uighur and Kazak word segmentation.

Keywords

Person name recognition, Uighur, Kazak, Rule-based method

1. Introduction

Name entity recognition (NER) is a interesting topic of information extraction, in which the major task is to identify and classify name entity from the text of natural languages. The name entity can be a word or word sequence of a person name, organization name, date, location, time, monetary value, or percentage expression. It is more specifically in the tasks of Information Extraction (IE) and Information Retrieval (IR). In this paper, we describe a rule-based Uighur person name recognition approach for Xinjiang Uighur language.

The Uighur and Kazak languages belong to the Altaic branch of Turkic languages. They are alphabetic languages, which are different from Chinese. The texts of this language are written from right to left, and lines are arranged from upper part to lower. Recognition of Uighur and Kazak person name is a very important problem in multilingual text processing.

Many methods have been used in name entity recognition for other languages. Some systems are based on statistical methods, such as Hidden Markov Models (HMM)[1,4]. Some are based on linguistic methods which make use of grammar rules[2,5]. Some methods are combinations of the rule-based and statistics approaches [3,6].

2. Related work

In general, research on English Named Entity (NE) identification has been focused on the machine learning method, such as Hidden Markov Models (HMM), decision tree and transformation-based learning, maximum entropy model, etc. Some methods have been applied to real application systems. Research on Chinese NE identification has to deal with the word segmentation problem. Therefore, there is usually a word segmentation process before NE identification.

We integrate word segmentation [11] and NE identification using a rule-based language model for Uighur and Kazak person name recogntition. Root-affix and syllable segmentation of Uighur and Kazak word are useful cues in text processing. There are various types of affix. They are linked to a root or another affix in different ways. The linkages are also complex. We propose methods for handling the basic name features of Uighur and Kazak words.

3. Background of person names

Uighur and Kazak person names (PN) are different from Chinese names. They are more like English and Arabic names. A Uighur and Kazak person name is formed in the pattern of FN+LN+(LN), in which there is a first name (FN), a last name(LN, father's name) and another optional last name (LN, grandfather's name). Most full person name consists of two parts: First name and last name, e.g.

Tuhti mehmut (in Uighur) توختى مەخمۇت

Here *Tuhti* is the first name, and it also can be used as his children's last name.

Here are some rules when people choose a name [8][9]:

- Use things from nature as person name.

  Polat (Steel, both in Uighur and Kazak) پولات
- Use a pioneer name from the Kuran.

  Yusup (both in Uighur and Kazak) يۇسۇپ
- Use words which mean good wishes.

  Tursun (Stop or remain, in Uighur) تۇرسۇن

  Kaynar (Source, in Kazak) قاينار
- Names attached with time attribute

  (conflict, in Uighur) كۈرەش

  (conflict, in Kazak) كۇرەس
- Use animal name as person name

  Burkit (eagle, in Kazak) بۇركت

4. Rule for person name identification

In this paper, we analyse identification rules for Uighur and Kazak names based on rules, and we will investigate research some methods for name identification. Name recognition involves the detection of their boundaries, the start and the end of all the possible spans of the name. Possible person names can be identified by looking at punctuation marks, title words, non-name entity, affix, etc. The rule can be summarized as follows:

4.1 Analysis of person names

In general, there are different proper names for male and female, and some words are used as not only person name but also common noun or adjective. For example,

- Female proper name: e.g.
  - Aygul（in Uighur) ئايگۇل
  - (in Kazak) ايگۇل
- Male proper name: e.g.
  - Azat Azatjan (in Uighur) ئازات، ئازاتجان
  - Azat Azatbek (in Kazak) ازات،ازاتبەك
- Both female and male proper name:
  - Ikbal Tuktax Nusiret (in Uighur)
    نۇسرەت ، توقتاش ، ئىقبال
  - Aray (in Kazak) اراي
- Common noun:
  - Guzel Yalkun (in Uighur) يالقۇن ، گۈزەل
  - Dulkun Marjan (in Kazak) دولقۇن ، مارجان

4.2 Limit element for PNR

(1) Title words

A person's title and appositive phrases are the mainly heuristic information for name identification, which helps us identify the candidates of person's name.

- Honorific title used for different ages. For example,
  - Ahun ئاخۇن
- Professional title.
  - (Teacher, in Uighur) مۇئەللىم
  - (Teacher, in Kazak) مۇعالىم
- Specialty title
  - Azat, (my grandfather, in Uighur) ئازات تاغام
  - Azat (my grandfather in Kazak) ازات اتام

A person's title may be placed in different positions in a sentence. For example,

- Placed in front of a person's name.
  - (Father or teacher, in Uighur) ئاكا، مۇئەللىم
  - (Father or teacher, in Kazak) اكە،مۇعالىم
- Use behind a person's name.
  - (Classmate, Friend, in Uighur) ساۋاقداش ، دوستۇم
  - (Classmate, Friend, in Kazak) ساباقتاس ،دوسىم
- Use in front of or behind a person's name.

(Comrade, in Uighur) يولداش
(Comrade, in Kazak) جولداس

(2) Added elements:

Some names of Uighur and Kazak have the following features: When the final vowel changes, the harmony of the vowel and the consonant, and syllable segmentation also change. This helps the word segmentation for Uighur and Kazak text. [10][11] The following is an example.

ئايگۈلگە = ئايگۈل + گە
قۇربانغا = قۇربان + غا

Some heuristic information is the constraint of the gender of person name candidates.

(Azatjan, in Uighur) ئازاتجان = ئازات + جان

(Azat,Azatbek, in Kazak) ازاتبەك = ازات + بەك

- Aygul (in Uighur) ئاي + گۈل = ئايگۈل
- “جان ”(in Uighur) and بەك (in Kazak) are often attached after a male name. This makes the name like a pet name.
- “گۈل . ”(which means flower, both in Uighur and Kazak) is often attached after a female name.

5. The experiment and results

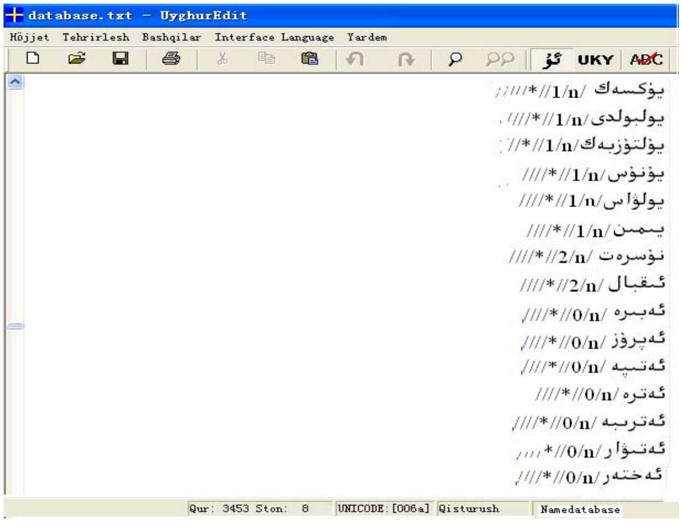
We have built a PN recognizer using a mix of rules, lexicon and some training strategies. Our system was implemented on Windows 2000 platform. “Ilikyurt Utm” input method was used to input the text. The system was programmed using Borland C++ with the database support of Microsoft Access 2000.

5.1 Person name lexicon

There are two lexicon used in the system. One is the common name lexicon and the other is the title lexicon.

The first lexicon contains 3453 the most commonly used Uighur person names, which are taken from a Uighur person name dictionary [7]. Each lexicon item contains name, gender, frequency, attribute, rule, etc.

The second lexicon contains different titles which are attached in front of or behind a name. The attribute values are the surface string from of the names. A heuristic partial matching method is implemented for person name identification. Each lexical item corresponds to a unique word of Uighur, each entry consists of a sequence of fields, the fields are terminated by vertical bars "|".


database.txt – UyghurEdit
Höjjet Tehrirlesh Bashqilar Interface Language Yardem
ئۇ UKY ABC

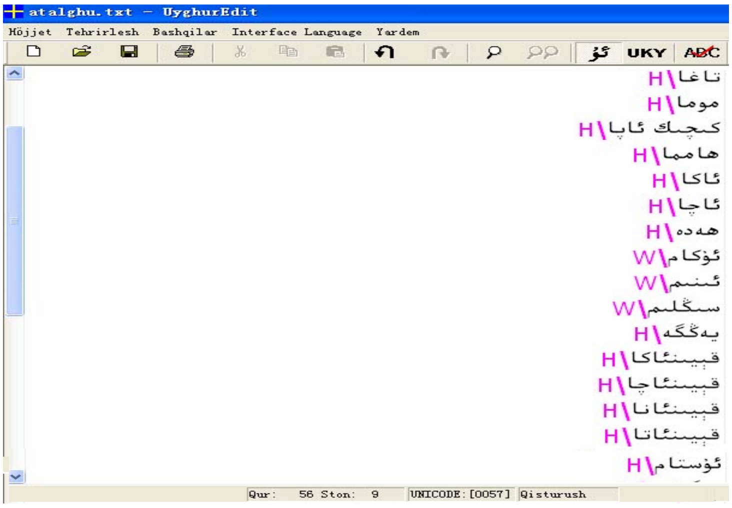
```
////*//1/n/ يۈكسەك
////*//1/n/يولبولدى
//*//1/n/يۈلتۈزبەك
////*//1/n/يۈنۈس
////*//1/n/يولۋاس
////*//1/n/يەمەن
////*//2/n/ نۇسرەت
////*//2/n/ ئىقبال
////*//0/n/ ئەبرە
////*//0/n/ ئەپرۇز
////*//0/n/ ئەتىپە
////*//0/n/ئەترە
////*//0/n/ ئەترىبە
,,,,*//0/n/ئەتىۋار
////*//0/n/ئەختەر
```

Qur: 3453 Ston: 8
UNICODE:[006a]
Qisturush
Namedatabase

Figure 1:Person-name lexicon

The following is a breakdown of the important fields in the person name dictionary file:

- *T*he headword of the lexical entry.
- *A* terminal symbol ('n' for Noun, 'a' for adjective).
- Gender attribute.
- Common noun attribute
- Frequency
- *T*itle information.
- added element used by the morphology .
- *P*ositions of the word in sentence.



```
H\تاغا
H\موما
H\كىچىك ئاپا
H\ھاما
H\ئاكا
H\ئاچا
H\ھەدە
W\ئۇكام
W\ئىنىم
W\سىڭلىم
H\يەڭگە
H\قېيىنئاكا
H\قېيىنئاچا
H\قېيىنئانا
H\قېيىنئاتا
H\ئۇستام
```

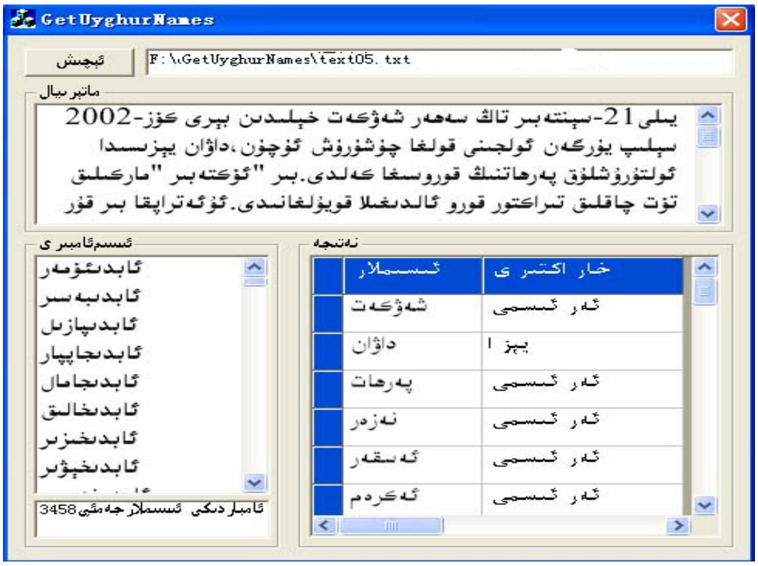
Figure 2 :Person-name title lexicon

5.2 Algorithm

Step1: Checking nouns and adjectives: nouns and adjectives are the two part of speech that are more likely to be part of a person's name.
Step2: Checking the Person's name lexicon for Uighur and Kazak name. It may be a person's name candidates
Step3: Checking the title or other rules for Uighur and Kazak name.
Step4: Rmoving the inflectional suffixes of Uighur and Kazak nouns and adjectives. Uighur and Kazak language are inflected languages. So nouns and adjectives have different forms in different cases.

5.3 Result and evaluation

After these processing stages, a series of evaluations have been done on the system. The training text corpus contains text from "Journal of Xinjiang Higher Education".



F:\\GetUyghurNames\text05.txt

يىلى 21-سېنتەبىر تاڭ سەھەر شەۋكەت خېلىدىن بېرى كۆز-2002
سېلىپ يۈرگەن ئولجىنى قولغا چۈشۈرۈش ئۈچۈن،داۋان يېزىسىدا
ئولتۇرۇشلۇق پەرھاتنىڭ قورۇسىغا كەلدى.بىر "ئۆكتەبىر" ماركىلىق
تۆت چاقلىق تىراكتور قورۇ ئالدىغىلا قويۇلغانىدى.ئۆئەتراپىقا بىر قۇر

ئابدۇئۆمەر
ئابدۇبەسىر
ئابدۇپازىل
ئابدۇجاپپار
ئابدۇجامال
ئابدۇخالىق
ئابدۇخىزىر
ئابدۇخېۋىر

ئامبىردىكى ئىسىملار جەمئىي3458

ئىسىملار	خاراكتېرى
شەۋكەت	ئەر ئىسمى
داۋان	يېزا
پەرھات	ئەر ئىسمى
نەزەر	ئەر ئىسمى
ئەسقەر	ئەر ئىسمى
ئەكرەم	ئەر ئىسمى

Figure 3 :Person-name recognition system

In order to evaluate the performance of our system, we calculated precision(P) and recall(R).

P=(number of correctly identified NE*)/(number of identified NE).*

R=(number of correctly identified *NE)/(number of all NE)*

Finally, it is calculated that the precision is 75% and recall scores is 80%.

In the case of person's names, a last name is made up of the father's name, this kind of strict case identification ensures a very high precision when matching last name for the person name lexicons. Also, the female's name didn't useful last name and some heuristic information improves the precision, recall of our system.

6. Conclusion

In this paper we have discussed person-name recognition for Xinjiang Uighur and Kazak in Xinjiang. It's based on linguistic methods which make use of name rules , some rules of Uighur and Kazak word segmentation. We tested our system using text from Journal of Xinjiang Institutions of Higher Learning. In future, we will be focused on location name and organization name recognition for Uighur and Kazak text.

Acknowledgement

This work is supported by the national Natural Science Foundation of China under Grant No. 69963002; Ministry of Education of the People's Republic of China under Grant No. MZ115-022.

References

1. Daniel M.Bikel, Miller S., Schwartz R. and R. Weischedel ,1997, Nymble: a high-performance learning name-finder. Proc. Of the 5th Conference on Applied Natural Language Processing,.
2. Dimitra Farmakiotou, etc., 2000 Rule-based named entity recognition for Greek financial texts. Proceedings of the International Conference on Computational Lexicography and Multimedia Dictionaries COMLEX 2000.
3. Mikheev A., etc,1998, Description of the LTG system used for MUC-7. Proceeding of Message Understanding conference(MUC-7).
4. 黄德根，王省，杨元生,2001,基于统计方法的中文姓名识别,中文信息学报, 15(2), PP31-37.
5. 孙茂松，黄昌宁，高海燕，1995,中文姓名的自动识别,中文信息学报,9(2),PP16-27.
6. 王省，黄德根，杨元生，1999，“基于统计和规则相结合的中文姓名识别”，《计算语言学文集》，黄昌宁，董振东主编，清华大学出版社，北京，PP155-161.
7. 木太里夫・司地克,1996，《维吾尔人名手册》，巴州文体局出版。
8. 穆台力甫 •司地克,1994, 维吾尔族人名和姓名初探,喀什师范学报,第3期PP89-95。
9. 木哈西，1988，《哈萨克人名及其写法》，伊犁人民出版社。
10. 哈米提・铁木尔，1987，现代维吾尔语语法，民族出版社。
11. 古丽拉・阿东别克，米吉提・阿布力米提，2004，维吾尔语词切分方法初探，中文信息学报,18（6）,PP31-37.