Person Name Recognition for Uyghur Using Conditional Random Fields

Muhtar Arkin[1], Abdurahim Mahmut [2], Askar Hamdulla[3*]

[1]College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang, 830046, P.R. China

[2]College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang, 830046, P.R. China

[3]College of Software, Xinjiang University, Urumqi Xinjiang, 830046, P.R. China

Abstract

This paper describes the person name recognition system for Uyghur, a highly agglutinative language, using the conditional random fields (CRFs) approach. In this paper, our experiments with various feature combinations for Uyghur have been explained. We also described a method to build Uyghur corpus from a set of hand annotated sentences. Feature selection is an important factor in recognition of person names using CRF, we used features as like Context Words, Stems of words, Suffix and its length, whether a suffix is exist, first and last syllable of the word, POS Information, Dictionary feature etc. For evaluation, we perform several experiments using different feature settings. This model proved to have a Recall of 81.86%, Precision of 88.79% and F-score of 85.19%.

Keywords: *NER, Uyghur language, person name recognition, CRF, feature*

1. Introduction

Named entity recognition (NER) is the task of identifying and classifying tokens in an open-domain text document into predefined set of classes such as person, organization, location, company etc. NER also deals about the identification of time, date, currency etc. In short the objective of NER is to identify and classify every word or term in a document into some predefined categories of identifiers. Many research works have been conducted to prove the importance of NER to the other natural language processing tasks such as part of speech tagging, information retrieval, question answering, summarization, identification of multiword expression, machine translation etc. Since person names takes an important role in these application, recognizing of person names is also an important task.

This paper describes the work on recognizing the most important named entities that is person names for the Uyghur language. We have adopted the CRFs-based approach to recognize person names from Uyghur texts. Uyghur is a type of highly agglutinative language in nature and belongs to the Turkish language family of the Altaic language system, it is an alphabetical language, in which words are formed by affixes attaching to the stem (or root), due to the property of highly agglutinative, the affixes of the word in Uyghur plays the most important role in the structure of the language. Uyghur uses Arabic-based script in which texts are written from right to left with some modifications; words are separated by blanks or other punctuation. As like English and Arabic person names Uyghur person names are formed in the pattern of first name, middle name and last name. Patterns of Uyghur person names clearly stated in [8].

In the current task, we first build a corpus based on a set of hand annotated sentences, and then we perform several experiments using various feature configurations.

Works of NERs can be found for many languages with different techniques. Some tried to apply rule base approach for finding NERs like in [1], [2] and [3]. Then the era of NERs using machine learning technique starts in [4], [5], [6] and [7].

However, Uyghur person name recognition has not been studied much yet due to the lack of Uyghur language processing resource and tools. Only a few research works done by using rule base approach for finding person names in Uyghur like in [8].

2. The challenges of Uyghur Person Name Recognition

1. The detection of the existing person name in a text which is a quite easy sub-task if we can use the capital letters as indicators to determine where the person name starts and where they ends. However, this is only possible when the capital letters are supported in the target language, which is not the case for the Uyghur language (Figure 1 shows the example of two words where only one of them is a person name and both of them start with the same character). The absence of capital letters in the Uyghur language is the main obstacle to obtain high performance in person name recognition.

(Ahmat) ئەخمەت

(Tomorrow) ئەتە

Figure 1: An example of illustrating the absence of capital letters in Uyghur

2. A lot of person names in Uyghur can appear in the dictionary with some other specific meanings. For example:

(meaning 1: Person name; meaning 2:hope) ئۈمىد

(meaning 1: person name; meaning 2:steel) پولات

Figure 2: different meanings of a Uyghur word

3. Uyghur is a highly inflectional language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. A Uyghur word can be seen as the following composition:

Word = prefix + stem +suffix1 + suffix2 + …

For example: ئەخمەتتىكى=ئەخمەت +تىكى

Both prefixes and suffixes are allowed to be combinations, and thus a word can have zero or more affixes.

4. Uyghur is a relatively free word order language. Thus person names can appear in any positions in the sentences making the person name task more difficult compared to others. For example:

منىڭ ئىسمىم ئەخمەت (My name is Ahmat)

ئەخمەت منىڭ ئاكام (Ahmat is my brother)

بۇ ئەخمەتنىڭ كىتابى (This is Ahmat's book)

Figure 3: person name appears in different position in sentences.

In the figure above, the person name "Ahmat" appeared different positions with different syntactic role such as predicate, subject and attribute respectively in the top down sentences given in that figure.

5. Uyghur is a resource-constrained language. Annotated corpus, name dictionaries, sophisticated morphological analyzers, POS taggers etc. are still under the process of developing, are not yet available for commercial use.

3. Conditional Random Fields

The concept of Conditional Random Field [9] is developed in order to calculate the conditional probabilities of values on other designated input nodes of undirected graphical models. CRF encodes a conditional probability distribution with a given set of features. It is an unsupervised approach where the system learns by giving some training and can be used for testing other texts. The conditional probability of a state sequence X=(x1, x2... x_T) given an observation sequence Y= (y1, y2... y_T) is calculated as:

$$P(Y \mid X) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t)\right) \qquad (1)$$

Where $f_k(yt\text{-}1,yt,X, t)$ is a feature function whose weight λ_k is a learnt weight associated with f_k and to be learned via training. The values of the feature functions may range between -∞ …+∞, but typically they are binary. Z_X is the normalization
Factor:

$$Z_X = \sum_y (\exp \sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, X, t)) \quad (2)$$

This is calculated in order to make the probability of all state sequences sum to 1. This is calculated as in Hidden Markov Model (HMM) and can be obtained efficiently by dynamic programming. Since CRF defines the conditional probability P (Y|X), the appropriate objective for parameter learning is to maximize the conditional likelihood of the state sequence or training data.

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) \quad (3)$$

Where, $\{(x^i, y^i)\}$ is the labeled training data.

Gaussian prior on the λ' s is used to regularize the training (i.e., smoothing). If λ ~ N (0, ρ2), the objective function becomes,

$$\sum_{i=1}^{N} \log P(y^i \mid x^i) - \sum_k \frac{\lambda_i^2}{2\rho^2} \quad (4)$$

The objective function is concave, so the λ's have a unique set of optimal values.

4. Person Name Recognition Using CRF

4.1CRF model of Uyghur Person Name Recognition

The CRF model of Uyghur person name recognition (Figure 4) consists of mainly data training and data testing. The following subsection will brief about each step of this CRF model we have used.

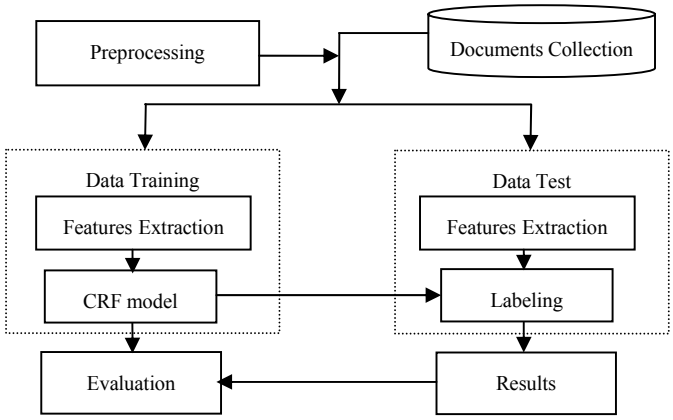

Figure 4： CRF model of Uyghur Person Name Recognition

4.2Training and test set collection

For the evaluation of Uyghur person name recognition, we have developed training and test data which has been manually tagged. This corpus has been developed from online and we have also refined the training and test data sets in order to remove the sentences which do not contain any named entity. After this refinement the training set consists of 4207 sentences which include 58058 words and test data contains 14581 words in 1051 sentences. As we mentioned above in section one, Uyghur person names like English, in the purpose of making it easy to deal with as like English person names, we carry out a conversion to our corpus, please see table 1 below for the conversion rule.

Table 1: A simple corpus conversion rules for Uyghur

No.	*Uyghur alphabet*	*Latin transcription*	*No.*	*Uyghur alphabet*	*Latin transcription*
1	ي	y	18	ھ	A
2	ا	a	19	ر	r
3	ل	l	20	ن	n
4	غ	G	21	ڭ	N
5	ۇ	u	22	چ	c
6	ز	z	23	ې	e
7	ك	k	24	ق	q
8	ش	x	25	خ	H
9	ى	i	26	ۇ	U

10	ت	t	27	ھ	h
11	و	o	28	گ	g
12	پ	p	29	ف	f
13	م	m	30	ۋ	w
14	س	s	31	ۆ	O
15	ب	b	32	ژ	J
16	د	d	33	ئ	v
17	ج	j			

4.3 Tagging Scheme

Each token of the corpus is tagged as below:

- B-PER: The beginning of the name of a person.
- I-PER: The inside of the name of a person.
- O: The word is not a person name entity.

4.4 Selection of Features

In order to get the best result, a carefully selection of feature is important in CRF. We have considered different combination from the following set for inspecting the best feature set for our task. The various features which are listed in our model are as follows:

- Context Word Feature. The previous and next words of a particular word are used as a feature. In our work we have experimented on different word window.
- Word Stem: Stemming is done to the words in the corpus and the stem of a word is used as a feature.
- Word Suffix. Word suffix information is helpful to identify person names. Suffix of the current and surrounding words are used as a feature.
- Suffix length: For every word the length of the suffixes is identified during stemming, if any and the length of suffixes is used as a feature.
- Scope of suffix length: Binary value '1' if suffix length is between 0-4, else '0'
- Binary notation if a suffix (es) is/are present: The suffixes play an important role in Uyghur since it is a highly agglutinative language. For every word if a suffix(es) is/are present during stemming a binary notation '1' is use otherwise'0'
- First and last syllable of the word: in Uyghur some of the first and last syllables are the part of person names, so we used this as a feature.
- POS Information. The first categorization of person names depends upon POS tag, because name entity is generally a combination of nouns. POS tagger is very helpful in tagging the data. POS feature is done manually for training data in our experiments. In the experiment we have used all the 39 POS tags.
- Dictionary feature: This list contains 8761 entries for the single person names. This feature is set to 1 for the current word if it is in the list otherwise it set to 0.
- Position of the current word in the sentence is also used as a feature.

4.5 Preprocessing and Feature Extraction

A Uyghur text document is used as an input file. The training and test files consist of multiple tokens. In addition, each token consists of multiple (but fixed number) columns where the columns are used by a template file. The template file gives the complete idea about the feature selection. Each token must be represented in one line, with the columns separated by tabular characters. A sequence of tokens becomes a sentence. Before undergoing training and testing in the CRF the input document is converted into a multiple token file with fixed columns and the template file allows the feature combination and selection.

An example sentences formation of few words in our model for feeding in the CRF tool is as follows:

yalGuz	a	0	0	null	0	0	0	yalGuz	yal	Guz	6	2	yENduK	1	O
kixilik	n	0	1	lik	1	1	3	kixi	ki	lik	7	3	yENduK	2	O
tiktak	n	0	0	null	0	0	0	tiktak	tik	tak	6	2	yENduK	3	O
top	q	0	0	null	0	0	0	top	top	top	3	1	yENduK	4	O
musabiKisidE	n	0	0	si	2	1	4	musabiKE	mu	dE	12	6	yENduK	5	O
^EkbEr	nr	1	0	null	0	0	0	^EkbEr	Ek	bEr	6	2	yENduK	6	B-PER
^ikkimiz	mr	0	1	miz	1	1	3	^ikki	ik	miz	8	3	yENduK	7	O
rEKibimizni	n	0	0	imiz	2	1	6	rEKib	rE	ni	11	5	yENduK	8	O
yENduK	v	0	0	N	2	1	4	yE	yEN	duK	6	2	null	9	O

Figure 5: An example sentences of data format

In the figure above, each column stands for the words, POS tag, dictionary feature, scope of suffix length, suffix of

words, number of suffix, whether suffix exist, suffix length, stem of words, first syllable, last syllable, length of word, number of syllable, near verb, position of the word in the sentence and person name tag respectively.

4.6Modeling

In order to obtain a model file we train with CRF using the training file. This model file is a ready-made file by the CRF tool for use in the testing process. In other words the model file is the learnt file after the training of CRF. We do not need to use the template file and training file again since the model file consists of the detail information of the template file and training file.

4.7 Testing

The test file is the test data into which we want to assign sequential tags of the Person Name Entity else 'O' for those words which are not Person Name Entity. This file has to be created in the same format as that of training file, i.e., of fixed number of columns with the same field as that of training file.

The output of the testing process is a new file with an extra column which is tagged with person name entity tagging scheme we used in part 4.3 of this section.

5. Experiments and Evaluation

Our experiments with CRFs were conducted using C++ based CRF++ 0.53 package [9]. Two types of features: unigram features and bigram features are used. We use standard measures: Precision, Recall and F1 to evaluate the performance of our recognition system. For measuring the performance of each experiment, we use the Perl script conlleval [10] provided by CoNLL-2000. Table 2 suggests the meaning of the notations used.

Table 2: Meaning of the notations

Notation	***Meaning***
W[-j,+j]	Context word features spanning from the j-th left position to the j-th right position
POS[-j,+j]	POS tag of the words spanning from the j-th left to the j-th right positions
Dw[-j,+j]	Dictionary feature of the words spanning from the j-th left to the j-th right positions
Sscope[-j,+j]	Suffix length scope feature spanning from the j-th left to the j-th right positions
Sufflen	Suffix length of the current word
hassuffix	Whether there is a suffix in the current word
suffnum	Number of the suffix of the current word
Suffix[-j,+j]	The suffix of the words spanning from the j-th left position to the j-th right position
Stem	Stem of the current word
Fsyll	First syllable of the current word
Lsyll	Last syllable of the current word

A lot of variations are noticed in Table 2 as we experiment with various feature combinations. We keep on changing the feature selection in order to get the highest possible Recall, Precision and the F-Score.

6. Conclusions

Our work concentrated on recognizing person names from open-domain text, first we have showed that Uyghur person name recognition meet with difficulties because of complicated characteristics of Uyghur and lack of resources, then we introduced a method to construct Uyghur corpus, experiments based on CRF model have shown that the efficiency of our approach. In our model feature selection is done through manual assumption but implementation of a Genetic Algorithm (GA) or other 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.

Table 3 Experimental Results

Feature	***Recall (in %)***	***Precision (in %)***	***F-Score (In %)***
W[-1,+1]	57.02%	94.98%	71.26%
W[-2,+2]	53.39%	94.66%	68.27%
W[-3,+3]	51.20%	94.56%	66.43%
POS[-1,+1]	79.33%	73.08%	76.07%
POS[-2,+2]	79.47%	73.71%	76.48%
POS[-3,+3]	79.12%	74.06%	76.51%
W[-1,+1], POS[-3,+3]	81.18%	86.13%	83.58%
W[-1,+1],POS[-3,+3],Dw[-2,2],Sscope[-2,+2]	81.31%	86.46%	83.81%
W[-1,+1],POS[-3,+3],Dw[-2,2],Sscope[-2,+2],Suffix[-1,+1],suffnum,hassuffix,Sufflen	81.31%	86.97%	84.05%
W[-1,+1],POS[-3,+3],Dw[-2,2],Sscope[-2,+2],Suffix[-1,+1],suffnum,hassuffix,Sufflen,stem	81.31%	87.80%	84.43%
W[-1,+1],POS[-3,+3],Dw[-2,2],Sscope[-2,+2],Suffix[-1,+1],suffnum,hassuffix,Sufflen,Stem,Fsyll[-1,+1],Lsyll[-1,+1]	81.52%	88.35%	84.80%
W[-1,+1],POS[-3,+3],Dw[-2,2],Sscope[-2,+2],Suffix[-1,+1],suffnum,hassuffix,Sufflen,Stem,Fsyll[-1,+1],Lsyll[-1,+1],Syllnum[-1,+1]	81.86%	88.79%	85.19%

Acknowledgements

This work was supported by Natural Science Foundation of China (61163033), Innovation Program for Excellent Ph.D. Candidates of Xinjiang University, Key Technologies R&D Program of China (2009BAH41B03), and Program for New Century Excellent Talents in University (NCET-10-0969).

References

[1] A. Mikheev, C. Grover and M. Moens, "Named Entity Recognition without Gazeteers", In the Proceedings of EACL, Bergen, Norway, 1999, pp.1–8

[2] Farmakiotou, V. Karkaletsis and S. K.J.S.G.S.C., "Rule-based Named Entity Recognition for Greek Financial Text", In the Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries(COMLEX 2000), 2000, pp.75–78

[3] H. Cunningham, "GATE, a General Architecture for Text Engineering", Computers and the Humanities, 36, 2002, pp.223–254

[4] G. Zhou and J. Su "Named Entity Recognition using an HMM-based Chunk Tagger", In the Proceedings of ACL, Philadelphia, 2002, pp.473–480

[5] D.M. Bikel, R.L. Schwartz and R.M Weischedel, "An Algorithm that earns What's in a Name", Machine Learning, 34, 1999, pp. 211–231

[6] K. Takeuchi and N. Collier, "Use of Support Vector Machines in Extended Named Entity Recognition", In the Proceedings of the 6th Conference on Natural language learning,(CoNLL-2002), 2002,pp.119–125

[7] H. Yamada, T. Kudo and Y. Matsumoto, "Japanese Named Entity extraction using Support Vector Machine", In the Transactions of IPSJ43, 2001, pp.44–53

[8] Gulila Altenbek, " Rule-based Person Name Recognition for Xinjiang Minority Languages", Journal of Chinese Language and Computing 15 (4): (219-226)

[9] http://crfpp.sourceforge.net/

[10] http://www.cnts.ua.ac.be/conll2000/chunking/output.htm

Muhtar Arkin received B.S in 2009 in Electronic Information and Science, from Xinjiang University of China. Currently a master student in the Institute of Information Science and Engineering of Xinjiang University.

Rahim Mahmut received B.E. in 1996 and M.E in 2009 in Electrical Engineering, from Xinjiang University of China. Currently, he is a lecturer in the Institute of Information Science and Engineering of Xinjiang University. He is pursuing his P.H.D in Natural language processing area under supervision of Professor Askar Hmadulla at Xinjiang University.

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, tutored by Professor Biing-Hwang(Fred) Juang. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 80 technical papers on speech synthesis, natural language processing and image processing. He is an affiliate member of IEEE.