UYGHUR MORPHOLOGICAL ANALYSIS USING JOINT CONDITIONAL RANDOM FIELDS: BASED ON SMALL SCALED CORPUS

GHALIP ABDUKERIM*

Xinjiang Technical Institute of Physical and Chemistry
Chinese Academy of Sciences, Urumqi 830011, China
and
University of Chinese Academy of Sciences
Beijing 100049, China

EZIZ TURSUN

Xinjiang Technical Institute of Physical and Chemistry
Chinese Academy of Sciences, Urumqi 830011, China
and
University of Chinese Academy of Sciences
Beijing 100049, China
and
Institute of Mathematics and Information of Hotan Teachers College
Hotan 848000, China

YATING YANG AND XIAO LI

Xinjiang Technical Institute of Physical and Chemistry
Chinese Academy of Sciences, Urumqi 830011, China
and
University of Chinese Academy of Sciences
Beijing 100049, China

ABSTRACT. As a fundamental research in the field of natural language processing, the Uyghur morphological analysis is used mainly to determine the part of speech (POS) and segmental morphemes (stem and affix) of a word in a given sentence, as well as to automatically annotate the grammatical function of the morphemes based on the context. It is necessary to provide various information for other tasks of natural language processing including syntactic analysis, machine translation, automatic summarization, and semantic analysis, etc. In order to increase the morphological analysis efficiency, this paper puts forward a hybrid approach to create a statistical model for Uyghur morphological tagging through a small-scale corpus. Experimental results show that this plan can obtain an overall accuracy of 92.58% with a limited training corpus.

1. Introduction. In Uyghur morphological analysis, the accuracy of the morpheme segmentation and tagging depends largely on the POS tagging of a given word. The transition model for morphological tagging puts forward in [14] based on the stem and affix dictionaries, which are used to obtain all possible segmentation

2010 Mathematics Subject Classification. Primary: 68T50, 68U15; Secondary: 60G60.
Key words and phrases. Uyghur language, natural language processing, morphological analysis, tagging corpus, Conditional Random Fields (CRF).

* Corresponding author: Ghalip Abdukerim.
candidates of each Uyghur word in a training corpus (set of words). Then a Markov model for the transition probability between the stem POS and affix functional tags is created with the corresponding morphological tag sequences of segmentation candidates as the training data before finally being used to determine the most likely segmentation result. This method can result in high accuracy primarily because the frequently used parts of speech, including nouns and verbs, hold a dominant position in the training and testing corpus. Although the results conform to the basic principles and rules of the statistical model, the determination of POS is “speculative” to some extent, especially when it comes to multi-class words without affixes due to the lack of context information between words in a sentence. Thus, it will influence accuracy and reliability of the statistical model. For example, the segmentation candidates of the word “باش (bash;)” are باش/N (noun: head), باش/D (adverb: initially), باش/A (adjective: primary) and باش/Q (quantifier: piece). If the POS is simply determined upon its frequency of use, without context, the word will always be treated as a noun. The absolute method is obviously unreasonable in a specific context.

Similarly, if functional tags of word morphemes are determined through morphological tag transition probability without other valuable context information, it is somewhat limited to choose the correct POS.

Owing to a complicated Uyghur morphological change and relatively high cost on large-scale manually tagged corpus, it is quite necessary to develop a small-scale manually tagged corpus to increase efficiency and reliability. This paper puts forward a hybrid approach that use a small-scale corpus as the training data in order to create a statistical model for the Uyghur morphological tagging and thus increase morphological analysis efficiency.

2. Review on linguistic features and related work.

2.1. Introduction to Uyghur language. Uyghur is a standard language commonly used among the Uyghur people. It is one of the main administrative languages in Xinjiang Uyghur Autonomous Region in China [13]. In Xinjiang Uyghur Autonomous Region, Uyghur is generally used in official activities, press and publication, ethnic minority education, social communication, social media technology, in addition to other fields. Approximately 13 million (2015) people speak Uyghur [2]. There is, though, few relevant research and delays are seen in the field of the corpus construction, corpus tagging system, and natural language processing due to the lack of the standard technology for Uyghur word processing. Some achievements and experience have been made and accumulated in the fields such as corpus construction, speech recognition, speech synthesis, statistical machine translation, and so on in view of the research results and technical means in relevant research fields both at home and abroad [4].

Uyghur belongs to the Turkic branch of the Altaic language family. It is an agglutinative language with a rich variety of morphologies. In Uyghur language, dozens of, or even hundreds of inflected forms of a word can be produced from a stem and a large number of inflectional suffixes to be affixed. One or more inflectional suffixes are affixed to the stem according to the grammatical rules of Uyghur and present a great deal of information in terms of the meaning, POS, number, case, tense, etc. Take the following Uyghur sentence as an example

ئاۋازەگەلىرىنى تاڭلىيىلىۋاتىسىم (I can’t hear your voices)
Formally, it consists of only two words, but the inflectional suffixes therein express extensive and complete information. Primarily:

On the one hand, the complicated structure of Uyghur leads to the vocabulary surge and subsequently gives rise to data shortage in statistical natural language processing, while on the other hand, it increases the difficulties in fields such as Uyghur syntactic analysis, machine translation, semantic analysis, automatic summarization, natural language understanding, and other relevant research fields.

Therefore, more in-depth research on Uyghur morphological change is of great significance to the improvement of main contribution in machine translation such as word alignment, POS tagging, named entity identification, Uyghur syntactic analysis, dependency parsing, and semantic analysis.

2.2. Related works. The objective of Morphological analysis is to correctly segment morphological components of a word and identify the category of corresponding forms. Morphological analysis depends largely on the grammatical and lexical features of a specific language. For example, the morphology of English is relatively simple. Additionally, the morphological analysis can be completely ignored regarding formation problems. Different forms of a word can be seen as other words. For instance, “opened, opens, and opening” can be seen as three different words. However, in such languages as Uyghur and Turkish, the morphological change cannot be handled in the same way as English because most words in these languages can have a number of agglutinative variations.

In recent years, some scholars have carried out related research on the linguistic features of Uyghur. The work in Tohuti [2004][9] systematically expounded the reduction, epenthesis, omission, of vowels and consonants occurring in the process of computer stemming. Murat Orhun, et. al. [2009][12] first conducted a research on Uyghur morphological analysis and segmented the stem and affix of an Uyghur noun through a rule-based method. However, no corresponding experimental result has been provided in relevant research reports.

The work in Wumaier et al. [2009][15] put forward the method of Uyghur noun stemming which combines the Conditional Random Fields (CRF) and the Finite State Machine in 2009. In this research, nouns from the Uyghur tagged corpus of Xinjiang University were used as both training and testing data. According to experimental results, the accuracy of the noun stemming reached 94%. The work in Aisha et al. [2009][1] carried out research on the Uyghur morphological statistical model based on the CRF in 2009. In that research, the stems and affixes manually segmented from the Uyghur words were used as both training and testing data. According to the final test results, the accuracy reached 89%.
The work in Ablimit et al. [2010][11], carried out a research on the Uyghur morphological segmentation based on a semi-supervised analysis with manually collected Uyghur stems, affixes, and compound affixes as testing data. Scholars trained the model through the forward-backward matching algorithm. According to experimental results, the accuracy for recognizing stem and suffix boundary reached 96% and the accuracy for the morpheme segmentation reached 92%.

The work in Xue et al. [2011][16] carried out a research through the rule-based method of morphological segmentation combined with semi-supervised training. According to experimental results, the accuracy, recall rate, and F value reached 81.4%, 72.3%, and 76.58% respectively.

The work in Aili et al. [2012][10] carried out research on Uyghur morphological structure with a directed graphical model. In this research, the nodes of the directed graph represented stems and affixes, as well as their respective POS and morpheme tags, whereas the edges of the graph represented transition. The manually tagged corpus provided by the multilingual key laboratory of Xinjiang University was used for training and testing. According to the experimental results, the stemming accuracy approximately reached 94% and the aligned F value for morpheme tagging reached 92.6%.

The research reported in Zhang et al. [2014][17] proposed a conjoint morphological analysis along with voice harmony restoration and morphological segmentation. This study solved the error propagation problem of the traditional morphological segmentation method. According to the results of the experiments based on the training and testing corpus provided by the multilingual key laboratory of Xinjiang University, the accuracy of the morphological segmentation increased by approximately 1.5% through the method presented in the research.

3. Semi-supervised Uyghur morphological analysis method based on hybrid approach. From the perspective of the natural language processing, the Uyghur morphological analysis is made to determine the POS and segmental morphemes (stem and affix) of a word in a given sentence and to facilitate the automatic annotation of the morphemes. It is based on the premise that the context provides the necessary information for other tasks of the natural language processing which may be syntactic analysis, machine translation, automatic summarization, and semantic analysis, etc. various tasks require different levels of information.

It can be seen in Figure 1 that the task of the Uyghur morphological analysis consists of two subtasks: segmentation and tagging (POS tagging and morpheme grammatical function tagging.) On the other hand, from the perspective of grammatical features, there is a close relationship between the POS of a word and affix tags namely, if the POS of a word is determined, it will provide necessary information for the segmentation and tagging of affixes; otherwise, affixes and their tags will provide necessary information for the determination of POS.

In the absence of a large-scale manually tagged corpus, it is extremely hard to carry out morphological analysis and obtain better results through the method of statistical learning due to the data sparseness caused by a rich variety of morphologies for Uyghur words. Therefore, this paper presents a hybrid approach that combines a small-scale tagged corpus, dictionaries, and rules. Generally speaking, one is supposed to initially use the method based on the stem and affix put forward in [14] to obtain the segmentation candidates of each word in a sentence. One may also obtain the filtering rules automatically from the manually tagged corpus.
Figure 1: The morphological analysis result and hierarchical relationship of a Uyghur sentence

Figure 2: The Architecture of a semi-supervised morphological analysis based on the hybrid approach

4. Morphological analysis based on joint CRF model.

4.1. Conditional random fields. The Conditional Random Fields (CRF)[8] is the Markov random field of random variable $Y$ conditioned on random variable $X$. CRF represents conditional distribution $P(Y|X)$ with undirected graphs. Since the
model is based on conditional distribution, explicit representation is not required for the dependency of variable \( x \).

The Linear Chain Conditional Random Fields is a simple CRF which can be used for sequence tagging. In the conditional probability model \( P(Y|X) \), \( Y \) is the output variable and represents the tag sequence; \( X \) is the input variable and represents the observational sequence to be tagged. The conditional probability model \( \hat{P}(Y|X) \) is obtained through the maximum likelihood or regularized maximum likelihood with the training data during learning. For the given input sequence \( x \), figure out the maximum output sequence \( y \) of the conditional probability \( \hat{P}(Y|X) \) during its prediction [7].

Assuming \( P(Y|X) \) is a linear chain CRF, the conditional probability with the value of random variable \( Y \) is \( y \) and can be expressed as below under the condition that the value of random variable \( X \) is \( x \):

\[
P(y|x) = \frac{1}{Z(x)} \exp \left( \sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)
\]

Wherein,

\[
Z(x) = \sum_y \exp \left( \sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i) \right)
\]

Wherein, \( t_k \) and \( s_l \) refer to the feature functions; \( \lambda_k \) and \( \mu_l \) refer to the corresponding weight values. \( Z(x) \) refers to a normalization factor and the summation is done based on all possible output sequences. Formulas (1) and (2) are the basic forms of linear chain CRF, representing the predicted conditional probability of the given input sequence \( x \) for output sequence \( y \). In Formulas (1) and (2), \( t_k \) is a feature function defined at the edge which is called as transition feature and depends on current and previous positions; \( s_l \) is a feature function defined at the node which is called as state feature and depends on the current position. \( t_k \) and \( s_l \) both depend on the position and are local feature functions. Generally, the value of feature functions \( t_k \) and \( s_l \) is 1 or 0. When the feature conditions correspond with each other, the value is 1 or 0. The CRF completely depends on the feature functions \( t_k \), \( s_l \) and corresponding values for weight \( \lambda_k \), \( \mu_l \)[7].

If the weight vector is expressed as \( w \),

\[
w = (w_1, w_2, \ldots, w_k)^T
\]

If the global feature vector is expressed as \( F(y, x) \),

\[
F(y, x) = (f_1(y, x), f_2(y, x), \ldots, f_k(y, x))^T
\]

The CRF can be expressed as the inner product of vector \( w \) and vector \( F(y, x) \):

\[
P_w(y|x) = \frac{\exp(w: F(y, x))}{Z_w(x)}
\]

Wherein,

\[
Z_w(x) = \sum_y \exp(w: F(y, x))
\]

The linear chain CRF is a Log Linear Model with a large number of features from input. In natural language processing, these features may include adjacent words, word pairs, prefixes, suffixes, uppercase and lowercase letters, filed-related characteristics, and semantic characteristics of words, etc.
4.2. CRF-based Uyghur morphological analysis statistical model. The goal of Morphological analysis is to identify the corresponding POS tagging sequence, morpheme segmentation of each word, and corresponding morphological tags for a given sentence. In traditional methods based on the flow line, the tasks above are completed step by step in accordance with the order of the POS tagging, morpheme segmentation, and morphological tagging.

Table 1: Feature Template of POS Tagging Model

| Features                        | Description                                      |
|---------------------------------|--------------------------------------------------|
| $w_{i-2}pos_i$, $w_{i-1}pos_i$, | Unary context features of the word               |
| $w_i$pos$_i$, $w_{i+1}pos_i$,   |                                                  |
| $w_{i+2}pos_i$                  |                                                  |
| $w_{i-2}w_{i-1}pos_i$, $w_{i-1}w_i$pos$_i$, | Binary context features of the word              |
| $w_iw_{i+1}pos_i$, $w_{i+1}w_{i+2}pos_i$, |                                                  |
| $w_{i-1}w_{i+1}pos_i$           |                                                  |
| $h_1(w_i)pos_i$, $h_2(w_i)pos_i$, | n characters selected from                        |
| $h_3(w_i)pos_i$,                | the beginning of the word                        |
| $h_4(w_i)pos_i$,                |                                                  |
| $h_5(w_i)pos_i$,                |                                                  |
| $t_1(w_i)pos_i$, $t_2(w_i)pos_i$, | n characters selected from                        |
| $t_3(w_i)pos_i$,                | the end of the word                              |
| $t_4(w_i)pos_i$, $t_5(w_i)pos_i$ |                                                  |
| $pos_{i-1}$pos$_i$             | POS tag transition feature                       |

With traditional methods based on the flow line, the error of the POS tagging is passed to morpheme segmentation, and the error of morpheme segmentation is passed to morphological tagging. The backward passing of errors may seriously affect the accuracy of follow-up tasks. The useful information from subsequent models cannot be used in previous tasks. Given that the candidates of each word for analysis have been determined, it is more appropriate to choose the correct result through a joint model compared to methods based on the flow line. However, owing to the complicated parameters and extremely difficult training of the joint model, this method will not be presented in this paper.

**POS tagging model:** POS tagging is the process of assigning a part-of-speech to each word in a given sentence. In this paper, Linear Chain CRFs are used to complete POS tagging. CRF selects features from the standpoint of words and characters in the formulation of the feature template. Namely, the first group of features contains the contextual information of a word; the second features consists of the character strings at the beginning and the end of the word; the third group is the transition features. The feature template of the POS tagging CRF model is as shown in Table 1. In Table 1, $w_i$ is No.i word in the sentence; $pos_i$ is the part-of-speech of $w_i$; $h_n$ is the No.n letter counting from the beginning of the word $w_i$; $t_n$ is the No.n letter counting from the end of the word $w_i$.

**Morphological tagging model:** Morphological tagging includes morpheme segmentation and the tagging of corresponding grammatical functions. For Uyghur, if the tagging is done merely through the method of statistical learning, due to
agglutination, rich and varied affixes and phonetic changes, etc., it will make high
demands on the training corpus and appear slightly unrealistic. Therefore, the stem
and affix dictionaries were used in this research to obtain morpheme/morphological
tag sequence candidates in order to select correct analysis results. In the case of
morphological tagging, the modeling training can be carried out in the same way
as the POS tagging, i.e. by linear chain CRF. An instance of training corpus for
morphological tagging consists of the segmented morpheme sequences of a word and
corresponding tag sequences, equivalent to each word and corresponding POS tag
in the training corpus for POS tagging. The number of morphemes composing a
word is much less than the number of words composing a sentence. For this reason,
the features of the morphological tagging model are relatively fewer and the feature
template is simpler. The feature template of the morphological tagging model is
as shown in Table 2. In Table 2, $m_i$ is No.i morpheme in a word, and $t_i$ is the
grammatical function tag of $m_i$.

4.3. Joint decoding. The CRF prediction problem is to figure out the output
sequence (tag sequence) $y^*$ with the maximum conditional probability under the
condition of the given CRF $P(Y|X)$ and input sequence (observation sequence) $x$.
This means that it tags the observation sequence. The CRF prediction adopts the
well-known Viterbi algorithm[3].

The following expression can be derived from the Formula (5):

$$ y^* = \arg \max_y P_w (y|x) $$

$$ = \arg \max_y \frac{\exp (w.F(y,x))}{Z_w(x)} $$

$$ = \arg \max_y \exp (w.F(y,x)) $$

$$ = \arg \max_y w.F(y,x) $$

Therefore, the CRF prediction turns out to be the solution to the optimal path
for the maximum denormalization probability

$$ \max_y (w.F(y,x)) $$

(8)

In order to solve the optimal path, Formula (8) is expressed as below:

$$ \max_y \sum_{i=1}^{n} w.F_i (y_{i-1}, y_i, x) $$

(9)

Wherein,

$$ F_i (y_{i-1}, y_i, x) = (f_1 (y, x), f_2 (y, x), \cdots, f_K (y, x))^T $$

(10)

| Features | Description                      |
|----------|---------------------------------|
| $m_{i-2} t_i, m_{i-1} t_i, m_i t_i, m_{i+1} t_i, m_{i+2} t_i$ | Unary context features of the morpheme |
| $m_{i-2} m_{i-1} t_i, m_{i-1} m_i t_i, m_i m_{i+1} t_i$ | Binary context features of the morpheme |
| $m_{i+1} m_{i+2} t_i, m_{i-1} m_{i+1} t_i$ | Morphological tag transition feature |
The CRF-based Uyghur morphological analysis model presented in this paper consists of the POS tagging model and morphological tagging model. Both models are trained separately with manually tagged corpora. The biggest benefit of such training is that the model and parameter evaluation is relatively simpler. The next task is to select a correct result from the analysis candidates of each word in a given sentence. In the method of dictionary-based morpheme segmentation, if no constraint or voice harmony restoration rule is included, the segmentation candidates of each word will include all of the combinations of the morphemes and corresponding morphological tags. Hence the formula for solving the optimal path based on joint decoding can be expressed as below:

$$\max_{y} \sum_{i=1}^{n} (w_{1}F_{1}(y_{i-1}^{1}, y_{1}^{1}, x^{1}) + (w_{2}F_{2}(y_{i-1}^{2}, y_{1}^{2}, x^{2})) \tag{11}$$

Wherein, $w_{1}F_{1}(y_{i-1}^{1}, y_{1}^{1}, x^{1})$ refers to the corresponding POS tagging model; and $w_{2}F_{2}(y_{i-1}^{2}, y_{1}^{2}, x^{2})$ refers to the corresponding morphological tagging model. The junction point of the two sub-models is $y_{0}^{2}=y_{i}^{1}$, indicating that when the POS value of No.i word is $y_{i}^{1}$, one must select all segmentation candidates with the morphological tagging sequence beginning with $y_{i}^{1}$ and conduct the scoring with Formula (11) in order to find the optimal path. For ease of understanding, Table 3 shows the list of morpheme segmentation candidates for the words in the example sentence in Figure 1.

![Figure 3: Morphological Tag Decoding Process of Words in the Sentence](image)

Figure 3 shows the morphological tag decoding process of the sentence in Table 3. For ease of graphing, the number of each segmentation candidate is used to represent the node.

5. Experiment and analysis.

5.1. Experimental data and baseline system. Due to the lack of public Uyghur morphological tagged corpora, 1400 sentences were selected randomly from the fields of news and literature works and manually tagged in this research. The tagged content includes the POS of each word, stem and affix segmentation, and primary and secondary grammatical functions of affixes. Table 4 shows the contents of the manually tagged corpus.

Conventionally, this paper divides the tagged corpus into three parts: the training set, development set, and test set. Table 5 shows the details of the experimental data.

The experiment uses CRFsuite tools for training, adopts an Average Perceptron algorithm, and optimizes parameters by means of cross validation for the training data.
Table 3: List of Morphological Tag Candidates of Words in the Sentence

| Word No. | Word segmentation candidates sequence No. | Morpheme (from right to left) | Tag (from left to right) |
|----------|-------------------------------------------|------------------------------|--------------------------|
| 1        | تۇئۇر   1.1 N 1.2 نۇئۇر P 1.3 تۇئۇر P 1.4 تۇئۇز V 1.5 نۇئۇز N VERB_VOICE |
| 2        | دﯦﮭﻘﺎنﭽﯩﻠﯩﻖ 2.1 N 2.2 دﯦﮭﻘﺎنﭽﯩﻠﯩﻖ N CASE VERB GER 2.3 دﯦﮭﻘﺎنﭽﯩﻠﯩﻖ N NUM_SUF VERB GER |
| 3        | مە tatto 3.1 مەكەني 3.2 مەتەکە 3.3 مەتەکە N PERS CASE 3.4 مەتەکە N PERS VERB_AUX VERB_TENSE |
| 4        | نئۇئۇر 4.1 نئۇئۇز 4.2 نئۇئۇز N CASE NUM_CAT CASE |
| 5        | نئۇئۇز 5.1 نئۇئۇز N CASE NUM_CAT CASE 5.2 نئۇئۇز N CASE NUM_CAT CASE |
| 6        | کئۇرى 6.1 کئۇرى 6.2 کئۇرى V VERB_TENSE PERSON |
| 7        | . 7.1 . Y |

Table 4: Manually Tagged Corpus Format and Content Example

|          | p Christophenka | p Christophenka+ CASE | N+PC |
|----------|-----------------|-----------------------|------|
| نارئىسنى | A               | A                     | A    |
| نارئىسنى | N               | N                     | N    |
| نارئىسنى | N               | N+NUM_CAT+PERS        | N+PL+3SGPL |

Table 5: Details of Experimental Data

|          | Number of sentences | Number of words (including punctuation marks) | Number of Uyghur words |
|----------|---------------------|-----------------------------------------------|------------------------|
| Training set | 1000               | 12433                                         | 10391                  |
| Development set | 200              | 2564                                          | 2151                   |
| Test set   | 200                | 2492                                          | 2075                   |
Table 6: Experimental Results

| Method                          | Accuracy (%) |
|--------------------------------|--------------|
|                               | Stemming   | Morpheme segmentation | POS      | Overall  |
| Tag sequence Markov model     | 90.18       | 83.25                 | 86.17    | 75.13    |
| Joint CRF model               | **91.98**   | **85.79**             | **92.7** | **77.95**|
| Tag sequence Markov model, α=0.95 | **92.65**   | **88.47**             | **88.12**| **79.65**|
| Joint CRF model, α=0.9         | **92.85**   | **89.76**             | **92.6** | **80.73**|

As a result of the complicated output results of the Uyghur morphological analysis in this research, as well as the lack of public morphological tagged corpora, there is no comparison between the experimental results and research results described in references [10] and [6]. Therefore, this paper takes the experimental results based on morphological tagging Markov model discussed in [14] as the baseline system and carries out a comparative analysis for the experimental results. The performance evaluation still adopts the standard used in [14], i.e. the stemming accuracy, morpheme segmentation accuracy, POS tagging accuracy, and overall accuracy. Table 6 shows the experimental results.

5.2. Model performance optimization. Morpheme number parameter: Just as the parameter “Number of heuristic morpheme units” is used to enhance the performance of morphological tagging Markov model in [14], the parameter can influence the performance of the joint CRF model as well. Namely, the parameter “Morpheme number” is brought into the morphological tagging model.

\[
\max_y Score_{joint \ CRF} = Score_{POS \ CRF} + Score_{morpheme \ CRF} \tag{12}
\]

Wherein,

\[
Score_{POS \ CRF} = w^1 F^1_i (y^1_i, y^1_{i-1}, x^1_i)
\]

\[
Score_{morpheme} = \alpha \left( 1 - \frac{M}{M^}\right) + (1 - \alpha) \left( w^2 F^2_i (y^2_i, y^2_{i-1}, x^2_i) \right) \tag{14}
\]

\(\alpha\) refers to the linear combination parameter used to determine the “number of morpheme units,” \(M\) refers to the number of tags (morpheme units) in the current segmentation candidates. \(\frac{M}{M^}\) refers to the normalization constant, i.e. the total of morpheme units in all segmentation candidates of the current word. In Table 6, Line 3 and 4 show the experimental results after the inserting of the parameter “Number of heuristic morpheme units.” In the table, \(\alpha\) is the value which is determined through the parameter tuning of the development set and allows the model to have maximum performance.

Linear combination parameter of joint CRF: Since the joint CRF model presented in this paper is a linear combination of the POS tagging model and morphological tagging model, the two self-model weights are set in this experiment through parameters. Namely:

\[
\max_y Score_{joint \ CRF} = \beta Score_{POS \ CRF} + (1 - \beta) Score_{morpheme \ CRF} \tag{15}
\]

Figure 4 shows the relationship between the linear combination parameter obtained from the experiment and its overall accuracy.
5.3. **Feedback in filtering rules.** Dictionary-based morphological segmentation involves the enumeration and combination of all possible morpheme segmentation candidates and morphological tags. Such segmentation candidates include many candidates of which do not comply with the Uyghur grammatical rules at all. The improper segmentation not only slows down the processing, but also results in noise information which in turn decreases the accuracy. One must develop filtering rules and filter out the improper segmentation candidates. This will be very helpful for improving the system performance. [14] adopts the method of active learning and improves the model performance by means of manual observation and collection of filtering rules from the development set. This paper adopts a method to train the corpus to learn filter rules automatically and thus filter unnecessary information out. Actually, the connection between the stem and affix is completed based on the precedence of the category of grammatical functions to which these morphemes belong. This makes it possible to judge whether the segmentation is correct or not through the morphological tag sequence. Therefore, one must first add the beginning tag and ending tag before and after each tag sequence in the training corpus, e.g. `<s> N </s>, <s> N CASE </N>` ...Next, collect unduplicated ternary sets from the list:

\[ R_1 : < t_{i}, t_{i+1}, t_{i+2} > \]

Compose the binary group with the POS of the stem and morphological tags of the last suffix

\[ R_2 : < t_{0}, t_{-1} > \]

During the testing phase, obtain the segmentation candidates of the given word and retain the segmentation candidates which comply with the following rules:
Table 7: Analysis for the Influence of Filtering Rules on Morphological Tagging

| Method                                                                 | Accuracy (%) |     |     |     |
|------------------------------------------------------------------------|--------------|-----|-----|-----|
| (Joint CRF model, α=0.9, β=0.1)                                        |              |     |     |     |
| When filtering rules are not used                                      | 92.85        | 89.76 | 92.6 | 80.73     |
| Joint CRF model, α=0.9, β=0.1, When filtering rules are used           | 97.4         | 94.58 | 96.35 | 92.58     |
| Tag sequence transition model, α=0.95, When filtering rules are used   | 94.35        | 93.22 | 94.78 | 91.81     |

Valid(tags) = \[
\begin{align*}
(\forall T r i g r a m \subseteq tags) & \land Tr i g r a m \in R_1, length \leq 3 \\
(\forall T r i g r a m \subseteq tags) & \land Tr i g r a m \in R_1 \land tags \in R_2, length > 3
\end{align*}
\] (16)

Primarily, when the morpheme number of the segmentation candidates is or less than 3, the corresponding tag sequence should meet R_1; when the morpheme number exceeds 3, the corresponding tag sequence should meet R_1 and R_2 at the same time.

Table 7 shows the experimental results after the filtering rules are both used and not used when Parameter α, β remains unchanged (α=0.9, β=0.1). For the purposes of comparison, the results of the experiment adopting filtering rules based on the tag sequence Markov model are listed on the last line.

It can be seen from Table 7 that filtering out tag sequences which do not comply with grammatical rules can greatly increase the morphological analysis efficiency. It should be noted that although filtering rules can reduce segmentation candidates, they cannot eliminate the ambiguity. The segmentation ambiguity cannot be eliminated without a statistical model.

6. Conclusions. This paper presents a hybrid approach which combines a small-scale tagged corpus, dictionaries, and rules for Uyghur morphological analysis. Firstly, selecting the segmentation candidates of each word in a sentence based on the stem and affix. Meanwhile, setting filtering rules automatically from the manually tagged corpus and to train the POS tagging model and morphological tagging model. Finally, the two CRF models mentioned above are used for scoring and the most likely analysis of each word is chosen. The hybrid approach has an advantage that not only applicable for multiple tasks including POS tagging, morpheme segmentation, and morphological tagging (unnecessary information, if you really want to add this one, just write it in another sentence, otherwise it seems very long sentence) but also the overall accuracy of analysis result reaches 92.58% (97.40%, 94.58%, and 96.35% for stemming, morpheme segmentation, and POS accuracy respectively), which is better than other existing methods. It provides a powerful support for subsequent natural language processing tasks such as machine translation, automatic summarization, and syntactic analysis, etc.
REFERENCES

[1] B. Aisha and M. Sun, A statistical method for Uyghur tokenization, in International Conference on Natural Language Processing and Knowledge Engineering, (2009), 1–5.

[2] Uyghur Language, Available from: https://en.wikipedia.org/wiki/Uyghur_language.

[3] S. Dandapat, S. Sarkar and A. Basu, Automatic part-of-speech tagging for bengali: An approach for morphologically rich languages in a poor resource scenario, in ACL 2007, Proceedings of the Meeting of the Association for Computational Linguistics, June 23–30, 2007, Prague, Czech Republic, 2007.

[4] T. Ibrahim, B. Yuan, A survey on minority language information processing research and application in xinjiang, Journal of Chinese Information Processing, 6 (2011), 149–156.

[5] T. Klymchuk, Regularizing algorithm for mixed matrix pencils, Applied Mathematics and Nonlinear Sciences, 2 (2017), 123–130.

[6] O. Kohonen, S. Virpioja, L. Leppanen and K. Lagus, Semi-supervised extensions to morfessor baseline, Proceedings of the Morpho Challenge 2010 Workshop, 2010.

[7] T. Kudo, K. Yamamoto and Y. Matsumoto, Applying conditional random fields to japanese morphological analysis, in Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain, 6 (2004), 230–237.

[8] Lafferty, D. John, McCallum, Andrew, Pereira and C. N. Fernando, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[9] T. Litip, The possibility of handling phonetic harmony by computer in Uyghur, Journal of the Central University for Nationalities, 5 (2004), 108–113.

[10] A. Mairehaba, W.-B. Jiang, Z.-Y. Wang, Y. Tuergen and Q. LIU, Directed graph model of Uyghur morphological analysis, Journal of Software, 12 (2012), 3115–3129.

[11] A. Mijit, N. Graham, M. Masato, M. Shinsuke, K. Tatsuya and H. Askar, Uyghur Morpheme-based Language Models and ASR, Ipsj Sig Notes, (2010), 581–584.

[12] M. Orhun, A. Ceyd Tantug and A. Esref, Rule Based Analysis of the Uyghur Nouns, International Journal on Asian Language Processing, 1 (2009), 33–44.

[13] L. Tohti, Modern Uyghur Reference Grammar, China Social Science Press, Beijing, 2012.

[14] E. Tursun, D. Ganguly, T. Osman, Y. Yating, G. Abdukerim, Z. Junlin and L. Qin, A semisupervised Tag-Transition-Based markovian model for Uyghur morphology analysis, ACM Transactions on Asian and Low-Resource Language Information Processing (TAL-LIP), 16 (2016), 8–23.

[15] A. Wumaier, T. Yibulayin, Z. Kadeer and S. Tian, Conditional random fields combined fsm stemming method for uyghur, in IEEE International Conference on Computer Science and Information Technology, (2009), 295–299.

[16] H. Xue, Y. Yang, T. Osman, X. Li and R. Zhang, Uyghur word segmentation using a combination of rules and statistics, Advances in information Sciences and Service Sciences(AISS), 3 (2011), 105–113.

[17] H. Zhang, Q. Cai, W. Jiang, Y. Lv and Q. Liu, Joint voice harmony restoration and morphological segmentation for morphology analysis, Journal of Chinese Information Processing, 6 (2014), 9–17.

[18] L. Zhu, Y. Pan and J. Wang, Affine transformation based ontology sparse vector learning algorithm, Applied Mathematics and Nonlinear Sciences, 2 (2017), 111–122.

Received June 2017; revised October 2017.

E-mail address: Ghalip@ms.xjb.ac.cn
E-mail address: Eziztursun@gmail.com
E-mail address: Yangyt@ms.xjb.ac.cn
E-mail address: Xiaoli@ms.xjb.ac.cn