北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2018.067

字符级的维吾尔语形态协同分析方法

吐尔洪·吾司曼[1,2,3]　杨雅婷[1,2,3]　艾孜孜·吐尔逊[4]　程力[1,2,3,†]

1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011; 2. 中国科学院大学，北京 100049; 3.新疆民族语音语言信息处理实验室，乌鲁木齐 830011; 4. 和田师范专科学校数学与信息学院，和田 848000;
† 通信作者, E-mail: chengli@ms.xjb.ac.cn

摘要　针对维吾尔语中构形词缀种类多、构形复杂以及发生音变现象等问题，提出一种基于字符级的维吾尔语形态协同分析方法。该方法最大的特点是同时进行维吾尔语的形态切分、形态标注以及音变还原，把词素边界、形态标记以及音变信息用一个复合标记描述，采用字符序列的标注方法进行训练。实验结果显示，形态切分、形态标注及音变还原正确率分别达到 95.86%, 92.39%和 99.70%，系统总体正确率达 91.84%。

关键词　维吾尔语；形态分析；协同分析

Collaborative Analysis of Uyghur Morphology Based on Character Level

Turghun Osman[1,2,3], YANG Yating[1,2,3], Eziz Tursun[4], CHENG Li[1,2,3,†]

1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi 830011; 2. University of Chinese Academy of Science, Beijing 100049; 3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011; 4. Institute of Mathematics and Information of Hotan Teachers College, Hotan 848000;
† Corresponding author, E-mail: chengli@ms.xjb.ac.cn

Abstract　The Uyghur language has various inflectional affixes, complex structures and phonetic changes. The authors propose a collaborative analysis method for Uyghur morphology at character level. It includes three procedures: morpheme segmentation, morphological annotation and reduction of phonetic changes. The main characteristics of this method is to use a composite tag to represent the morpheme boundaries, annotations and phonetic changes. In addition, character sequence annotation is used to train the model. Experimental results show that the accurency of morpheme segmentation, morphological annotation and reduction of phonetic reaches 95.86%, 92.39% and 99.70% respectively. The overall accuracy of the system reaches 91.84%.

Key words　Uyghur; morphological analysis; collaborative analysis

在自然语言处理中，语法结构的分析离不开对词汇形态学的分析，句子语义的分析也离不开对词汇语义的分析，因此形态分析是自然语言处理中的基础性问题，在机器翻译、信息检索和问答系统等领域具有广泛的应用前景[1]。维吾尔语是一种黏着性语言，丰富的构形词缀连接词干表

中国科学院“西部之光”基金(2017-XBZG-BR-001,2017-XBQNZ-A-005)、国家千人计划项目(Y32H251201)、国家自然科学基金(U1703133)和中国科学院青年创新促进会(2017472)资助
收稿日期: 2018–04–15; 修回日期: 2018–08–09; 网络出版时间: 2018-08-22 13:40:39

示数、格、时态等语法功能，例如，词干“ياز”(夏天)作为名词时，连接属格缀“نىڭ”可以得到“يازنىڭ”(夏天的)，连接时位格缀“دا”得到“يازدا”(在夏天)；“ياز”(写)作为动词时，连接过去式形动态缀“غان”得到“يازغان”(写的)，嵌套连接能源体缀“الا”、否定缀“ما”、现在–将来时缀“ي”及代词缀“مەن”，得到“يازالمايمەن”(我不会写)，因此维吾尔语单词通过构形方式衍生出新词。

自然语言形态分析研究始于1955年，宾夕法尼亚大学的Harris进行了英语词素边界识别方面的研究[2]，1970年，他进行了英语词素边界识别方面的研究，并提出基于前文的词素边界识别思想[3]；1994年Merialdo[4]利用隐马尔可夫模型(HMM)进行基于无监督训练的英语词性标注研究；2001年Goldsmith[5]提出最小描述长度准则(minimum description length)的形态分析方法，在最小描述长度的启发式算法中，较短的字符串更有可能被选中为词干；2005年Creutz等[6]提出最大自然估计算法的形态分析工具Morfessor，并于2007年用最大后验估计算法对它进行改进，在英语、芬兰语、土耳其语和阿拉伯语语料的测试结果为66.2%，66.4%，70.7%和68.1%[7]；2013年Ruokolainen等[8]基于条件随机场(CRF)算法进行少量标注语料的有监督形态切分研究，实验结果显示，有监督训练的效果明显高于无监督训练。

针对维吾尔语的形态分析研究始于1997年，玉素甫·艾白都拉等[9]首次进行维吾尔语词法分析的研究；2006年阿依克孜·卡德尔等[10]对维吾尔语名词及词缀进行语法形态学方面的分析；艾山·吾买尔2008年做了有限状态自动机和词典相结合的维吾尔语名词词干切分研究，其测试结果为91%[11]，2009年采用有限状态自动机和最大熵模型(MEM)的混合模型，解决词干切分中的歧义问题[12]；2011年薛化建等[13]在词缀库的基础上，提出一种无监督维吾尔语词切分方法，该方法在测试集上准确率达到80.4%。

以上研究主要采用基于规则的方法，这类方法最大的缺点在于收集到的规则无法涵盖所有的语言现象，积累规则需要较高的语言学方面的知识，因此后来的研究中基于统计的方法成为主流方向。麦热哈巴·艾力等[14]提出维吾尔语词法分析有向图模型，该模型把有向图的节点当作词干和词缀，图边表示节点之间的转移概率，针对维吾尔语的音变现象，提出词内字符对齐算法的自动还原模型，并用统计的方法解决词内音变现象，实验结果显示最终词干提取正确率为94.7%。张海波等[15]把音变还原问题结合在形态切分过程中，有效避免了串行模型中音变还原对形态切分的错误传播问题。米尔阿迪力江·麦麦提[16]用Morfessor模型在大规模网络文本上做实验，其词干提取准确率达到86.08%。Tursun等[17]结合词典及规则进行形态切分，得到维吾尔语形态标记马尔科夫模型。哈里旦木·阿布都克里木等[18]采用双向门限递归单元神经网络进行维吾尔语的形态切分研究，通过门限递归单元有效处理长距离依赖问题。Maimaiti等[19]用BI-LSTM-CRF模型进行词性标注实验，并且在实验中论证特征对标注模型的重要性。

维吾尔语形态分析的目标是词干、词缀切分以及为它们的语法功能进行自动标注，是维吾尔语自然语言处理研究的首要任务。一方面，维吾尔语作为形态复杂的黏着型语言，具有众多的构形词缀、丰富的构形规则、歧义的边界及词性以及复杂的音变现象等特点，使维吾尔语形态分析成为具有挑战性的研究。另一方面，作为语料资源相对缺乏的语言，没有相应的开源标注语料库，从而制约了相关研究的进一步深入。本文在文献[15]的基础上，提出基于字符层面的协同分析方法，把形态切分、形态标注及音变还原任务同时定义为字符序列的标注问题，从而有效降低数据稀疏和数据量少等对形态分析的影响。

1　维吾尔语形态分析

维吾尔语的词缀有构词词缀和构形词缀两种，构词词缀的数量较少而且构词规则较为固定，衍生出的新单词一般在词典里可以查到，因此构词词缀不在本文的研究范围内。构形词缀数量众多，以不同的组合方式连接到词干可以产生多种形态变化，衍生出新的单词，使词汇量剧增，并且在组合过程中会发生音变现象，当不同的构形词缀嵌套组合时，表达的意思更为复杂，表1给出维吾尔语形态切分的例子。

目前维吾尔语形态分析研究面临的挑战如下。

1)维吾尔语词缀众多，构形方式丰富。维吾尔语是形态复杂语言，总共有 300 多个构形词缀，分为 17 个大类，如人称、格、比较级、时态等，每一类表达的语法意义各不相同，构形规则也不同。

2)词素有歧义。维吾尔语单词的词性有歧义，这种现象在形容词和副词中比较常见，如单词“كۆپ”表示“多”的意思，修饰名词时具有形容词特性，而修饰动作具有副词特性。同样，词缀“ش”当动名词缀被构形的动词具有名词性质，而当共同态缀时表示动作共同完成。

3)词素边界有歧义。有些单词的形态切分有歧义，如单词“قسقسى”有两种形态切分形式“قسقسى”和“قسقسا+سى”，第一种形式是助词“总之”，第二种形式为形容词“短”，连接第三人称词缀表示“短的”。

4)切分有歧义。有时不同的单词经过构形后，派生的单词在书写形式上是一样的，如“بەر”(给)和“بار”(去)由状态副动词缀“پ”构形后变成“ببرىپ”，此时很难用规则判断其词干。

5)音变现象。维吾尔语在构形过程中要遵循语音和谐规则，当词干和词缀，词缀和词缀相互连接时，有可能发生脱离、弱化、增音等音变现象，见表 2。

通过以上分析可以发现，一个维吾尔语词干以不同的构形方式派生出不同的新单词，派生过程中人称、数以及时态等语法信息以词缀的形式表达，这种现象在机器翻译系统中导致词对齐效率的降低，增加未登录词数量，从而影响译文质量[20–21]。在信息检索过程中对内容进行形态切分，可以压缩倒排表的大小，而且检索结果可以覆盖到拥有相同词干的所有单词，因此在缩短系统相应时间的同时，还可以得到较高的查全率[22]。图 1 为维–汉统计机器翻译中词干切分之前(a)和词干切分之后(b)的双语句子的对齐结果，图 1(a)中中文单词“加强”与它对应的维吾尔文单词“كۈچەيتسە”没有对齐，而把维吾尔文单词“كۈچەيتسە”错误地与中文单词“金融业”对齐。词干提取之后解决了图 1(a)中的对齐错误，如图 1(b)所示。

表 1 维吾尔语形态切分例子

Table 1　Example of Uyghur morphology

单词	切分形式	说明	中文
كىيىم	كىيىم	名词	衣服
كىيىمىڭ	كىيىم+ىڭ	名词+第二人称单数	你的衣服
كىيىمنىڭنى	كىيىم+ىڭ+نى	名词+第二人称单数+宾格	把你的衣服
كىيىمڭلارنى	كىيىم +ڭلار+نى	名词+第二人称复数+宾格	把你们的衣服
ئوقۇ	ئوقۇ	动词	读
ئوقۇدۇق	ئوقۇ+د+ۇق	动词+过去式+第一人称复数	我们读了
ئوقۇدى	ئوقۇ+ د+ى	动词+过去式+第三人称单数	他读了
ئوقۇلمىدى	ئوقۇ+ل+ما+د+ى	动词+被动态+否定+过去式+第三人称单数	(它)没有被读
ئوقۇيالىغانلار	ئوقۇ+يالا+غان+لار	动词+能愿体+过去式形动态+复数	能读的人(复数)

表 2 维吾尔语音变现象

Table 2　Example of phonetic change in Uyghur morphology

单词	中文	切分形式	音变现象说明
قەلەمىم	我的笔	قەلەم+ىم	词干第 4 个字符“ە”被弱化成“ى”
يېزىپ	写	يازـ+ىپ	词干第 2 个字符“ا”被弱化成“ې”
ۋاقتى	他的时间	ۋاقىت+ى	词干第 4 个字符“ى”被脱落
بويىنى	他的脖子	بويۇن+ى	词干第 4 个字符“ۇ”被脱落
يۇيۇپ	洗	يۇ+ۇپ	词干“يۇ”和词缀“ۇپ”的连接位置出现增音字符“ي”
دېيىشمىدى	他没有说	دە+ش+مە+د+ى	词干第 2 个字符“ە”被弱化成“ې”；词干和词缀“ش”的连接位置出现增音字符“ي”；词缀“مە”第 2 个字符被弱化成“ى”

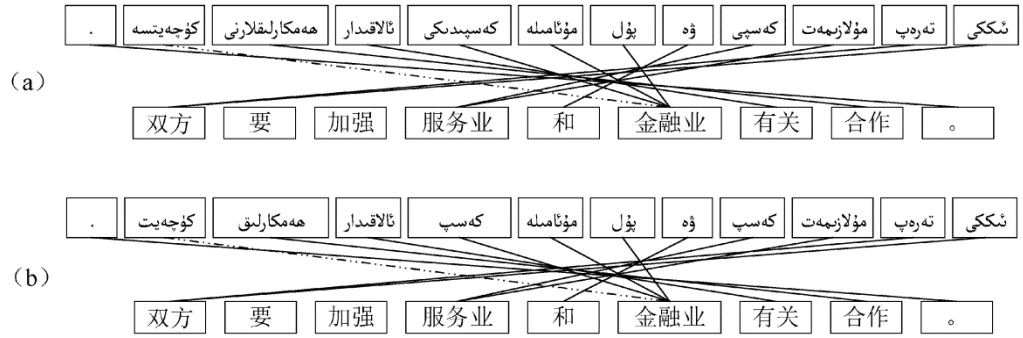

图 1　统计机器翻译中的词对齐

Fig.1　Word alignment in statistical machine translation

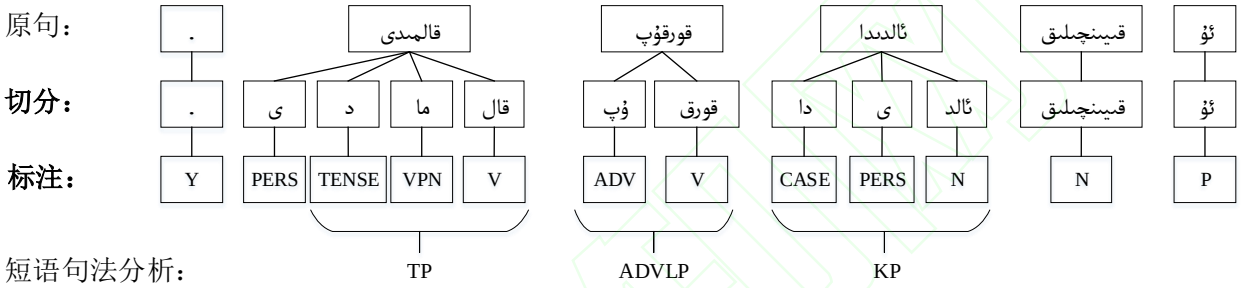

图 2　词法和句法分析过程

Fig. 2　Process of morphological analysis and syntactic parsing

维吾尔语形态分析目的包括形态切分、形态标注及音变还原，是进行句法分析研究的前提，对有效辨别短语结构提供帮助，如：　名词和代词与格词缀组成格短语 (kasus phrase, KP)，动词与时态词缀组成时态短语(tense phrase, TP)。图 2 显示例句“.ئۆ قىينىچلىق ئالدىدا قورقۇپ قالمىدى”(他在困难面前没有害怕。)从词法分析到句法分析的过程：　第 1 层是原句的分词状态，第 2 层是句子中各单词的词素切分状态，第 3 层是各词素的词性标注，第四层是其短语结构。

2　形态协同分析方法

2.1　任务定义

假设维吾尔语单词由字符序列$c=(c_1,c_2,c_3\dots c_n)$组成，其中 n 为单词的长度。$l=(l_1,l_2,l_3\dots l_n)$表示c的形态特征序列,其中l_i表示字符c_i的形态特征信息，包括所在词素的边界信息、词性信息以及所在位置的音变现象信息，可由l_i的上下文信息预测得到。因此本文把单词的形态分析任务定义为字符序列的标注问题，即通过观察序列c,得到其形态特征序列l。

Lafferty 等[23]提出把线性链条件随机场(linear chain)应用于标注问题的思路，条件随机场是一种概率无向图模型，由无向图表示联合概率分布，当给定随机变量 X 的条件下，预测随机变量 Y。本文中单词被划为若干个字符组成的字符序列，其条件随机场模型定义为

$$P(l|c)=\frac{1}{Z(c)}\exp\Big(\sum_{i,m}\lambda_m\,t_m(l_{i-1},l_i,c,i)+\sum_{i,n}\mu_n s_n(l_i,c,i)\Big),\tag{1}$$

其中，$Z(c)=\sum_y \exp\left(\sum_{i,m}\lambda_m t_m(l_{i-1},l_i,c,i)+\sum_{i,n}\mu_n s_n(l_i,c,i)\right)$为归一化因子，$t_m$和$s_n$是局部特征函数，都依赖于位置信息。通常特征函数$t_m$和$s_n$起开关作用，满足特征条件时取值为 1，否则为 0。条件随机场完全由特征函数t_m, s_n以及对应的权值λ_m, μ_n确定[24]。

2.2 模型描述

从维吾尔语单词词素边界识别的角度考虑，标记可以简单地设置为{B, I}标记。其中，B 代表词素的起始位置，I 代表词素的非起始位置。为了达到维吾尔语形态协同分析的目的，本文扩充{B, I}标记方式。具体来讲，把只有形态切分功能的标记扩充为同时包含形态切分，形态标记以及音变还原功能的标记。

词素切分由{B, I}标记表示。形态标记由词素的词性(POS)表示，如，“N”表示名词、“V”表示动词、“CASE”表示格词缀等。音变现象分别由{N, I, R, S}标记表示。N(none)表示没有发生音变现象。I(insert)为增音标记，表示当前字符在构形过程中被增加的字符。R(remove)为脱落标记，表示当前和下一个字符之间发生字符脱落的现象。R 标记连带一个字符位，表示被脱落的字符，如，“R_ى”表示字符“ى”发生脱落现象。S(substitute)为弱化标记，表示当前字符在组合过程中被弱化。S 标记同样连带一个字符位，表示被弱化的原始字符，如，“S_ا”表示字符“ا”被弱化成当前字符。下面以单词“دېيشمدى”(他们没有说)为例，具体的处理流程如下。

1)　语料库中所有单词进行人工切分得到对应的词素序列。单词“دېيشمدى”的词素切分形式为“دە+ش+مە+د+ى”。

2)　词素序列进行形态标注。词素序列“دە+ش+مە+د+ى”的形态标注序列为“V+VVOICE+VPN+TENSE+PERS”。

3)　通过字符对齐方法识别词素内的音变现象，得到对应的音变还原标记序列。音变还原过程由式(2)表示，其中i为原单词c的当前字符索引，j为词素序列m的当前字符索引。

$$f(c,m,i,j)=\begin{cases}\text{N, 当 } c_i=m_j,\\ \text{S, 当} c_i<>m_j \ \ and\ c_{i-1}=m_{j-1}\ and\ c_i\ in\ ['ې','ى']\\ \text{R, 当} c_i<>m_j \ \ and\ ((c_{i-1}=m_{j-1}\ \ or\ \ f(c,m,i-1,j-1)=S)\ and\ (c_i=m_{j+1}\ \ or\ \ f(c,m,i,j+1)=S))\\ \text{I, 当} c_i<>m_j \ \ and\ ((c_{i-1}=m_{j-1}\ \ or\ \ f(c,m,i-1,j-1)=S)\ and\ (c_{i+1}=m_j\ \ or\ \ f(c,m,i+1,j)=S))\end{cases} \tag{2}$$

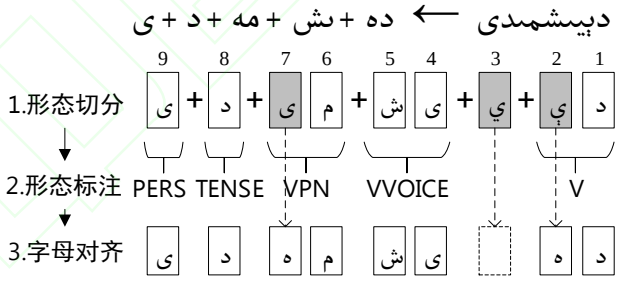

图 3　字符对齐图

Fig. 3　Letter alignment in Uyghur word-morpheme segmentation

字符对齐形式如图 3 所示，各个词素的起始位置分别在 1, 4, 6, 8 和 9 位，从图 3 可以发现，单词“دېيشمدى”的第 2 和第 7 字符“ە”分别弱化成字符“ې”和“ى”，得到“S”标记。第 3 个字符“ي”为增音字符，得到“I”标记。

4)　最终根据词素的形态标注信息、字母对齐信息以及音变还原信息，最终得到如表 3 所示的协同标记序列。

2.3 特征选取

在训练和解码过程中, 序列标注模型需要加入特征元素, 并且特征的优劣直接影响模型的预测能力。在字符序列中, 上下文关系是最重要的特征信息, 因此本文在上下文中分别取 1 个字符, 2 个字符和 3 个字符作为模型的特征。表 4 给出本文选取的特征模板, 其中 i 为观察窗口的半径 , 并且列举当前字符为单词“دىيىشمدى”的第 5 个字符, 观察窗口半径为 4 时的特征例子。

3 实验

3.1 实验数据

本文实验所用语料是人工进行形态标注的 3533 条句子, 包含政务新闻、法律法规以及文学类内容, 其中维吾尔语单词有 54039 条(词项 12700 条), 其中有 4564 条单词发生音变, 表 5 给出语料中各种音变现象的统计数据。从语料中抽取 90%作为训练数据(3180 条句子, 包含 48663 条单词), 其余 10%的句子作为测试数据(353 条句子, 包含 5376 条单词)。

表 3 维吾尔语形态标注符合标记

Table 3　Compound labels used in Uyghur morpheme segmentation

序号	字符	切分标记	形态标记	音变还原标记	协同标记
1	د	B	V	N	B_V_N
2	ې	I	V	S_ه	I_V_S_ه
3	ي	-	-	I_ي	I_V_I_ي
4	ى	B	VVOICE	N	B_VVOICE_N
5	ش	I	VVOICE	N	I_VVOICE_N
6	م	B	VPN	N	B_VPN_N
7	ى	I	VPN	S_ه	I_VPN_ S_ه
8	د	B	TENSE	N	B_TENSE_N
9	ى	B	PERS	N	B_PERS_N

表 4 特征模板

Table 4　Feature template

模板	说明	例子
$c_{-i}, c_{-(i-1)}, \ldots, c_{-1}, c_1, \ldots c_{i-1}, c_i$	一元特征	د، ې، ي، ى، م، ى، د، ى
$c_{-i}c_{-(i-1)}, \ldots c_{-2}c_{-1}, c_1c_2, \ldots c_{i-1}c_i$	二元特征	دې، يې، يي، ىي، مى، ىد، دى
$c_{-i}c_{-(i-i)}c_{-(i-2)}, \ldots c_{-3}c_{-2}c_{-1}, \ldots c_1c_2c_3, \ldots c_{i-2}c_{i-1}c_i$	三元特征	دېي، يېي، ىيې، مىد، ىدى

表 5 语料库中音变现象的统计情况

Table 5　Statistics of phonetic changes in the dataset

音变现象	数量	比重/%
音变单词	7116	13.17
弱化	6931	12.83
脱落	168	0.31
增音	46	0.08

表 6 评价指标

Table 6　Evaluating indicators

评价指标	公式	解释
词素边界识别率	$\mathrm{MBI}_{\mathrm{Accuracy}}=\frac{\text{正确切分词素的单词数}}{\text{单词总数}}\times 100\%$	正确切分词素的单词比例
词干提取识别率	$\mathrm{SS}_{\mathrm{Accuracy}}=\frac{\text{正确切分词干的单词数}}{\text{单词总数}}\times 100\%$	正确切分词干部分的单词比例
形态标注准确率	$\mathrm{MA}_{\mathrm{Accuracy}}=\frac{\text{正确切分词素的同时正确标注形态功能的单词数}}{\text{单词总数}}\times 100\%$	正确切分并正确标注的单词比例
音变还原准确率	$\mathrm{PR}_{\mathrm{Accuracy}}=\frac{\text{正确还原发生音变现象的单词数}}{\text{发生音变的单词总数}}\times 100\%$	正确识别音变现象的单词比例
准确率	$\mathrm{Accuracy}=\frac{\text{正确标注的单词数}}{\text{单词总数}}\times 100\%$	正确处理以上指标的单词比例

表 7 观察窗口半径取值不同情况下的实验结果(%)

Table 7　Experimental results based on different half window size (%)

观察窗口半径	2	3	4	5	6	7	8	9	10
$\mathrm{MBI}_{\mathrm{Accuracy}}$	93.75	95.75	96.35	96.39	96.37	96.39	96.31	96.31	96.27
$\mathrm{SS}_{\mathrm{Accuracy}}$	93.88	95.92	96.54	96.59	96.59	96.61	96.54	96.54	96.54
$\mathrm{MA}_{\mathrm{Accuracy}}$	89.37	91.68	92.54	92.78	92.67	92.76	92.72	92.70	92.68
$\mathrm{PR}_{\mathrm{Accuracy}}$	99.45	99.71	99.76	99.79	99.79	99.79	99.79	99.79	99.70
Accuracy	88.52	91.46	92.37	92.59	92.48	92.57	92.54	92.50	92.50

表 8　CRF、Morfessor、HMM 和 MEM 下的对比结果(%)

Table 8　Experimental results on CRF, Morfessor, HMM and MEM (%)

模型	CRF	Morfessor		HMM	MEM
		非监督	监督		
$\mathrm{MBI}_{\mathrm{Accuracy}}$	96.39	44.75	48.12	64.84	94.58
$\mathrm{SS}_{\mathrm{Accuracy}}$	96.59	51.00	57.79	66.33	94.88
$\mathrm{MA}_{\mathrm{Accuracy}}$	92.78	–		46.17	84.80
$\mathrm{PR}_{\mathrm{Accuracy}}$	99.79	–		98.06	99.68
Accuracy	92.60	–		45.65	84.56

3.2 实验结果

本文使用 CRFsuite[1]作为训练和解码工具。为了确定最优特征模板的窗口半径，在窗口半径为 2~10 的范围内进行 9 次实验。本实验在准确率(Accuracy)做评价指标的基础上，还使用词素边界识别率($\mathrm{MBI}_{\mathrm{Accuracy}}$)、词干提取[2]识别率($\mathrm{SS}_{\mathrm{Accuracy}}$)、形态标注准确率($\mathrm{MA}_{\mathrm{Accuracy}}$)以及音变还原准确率($\mathrm{PR}_{\mathrm{Accuracy}}$)等指标，表 6 列出各个评价指标的定义，表 7 给出窗口半径在取值不同时的实验结果。

在同样的训练集和测试集上，用 Morfessor[3]、HMM[4]和最大熵(MEM)[5]模型分别做 3 次实验，表 8 列出 CRF、Morfessor、HMM 及 MEM 模型下的最好实验结果。

3.3 实验分析

3.3.1 协同形态分析实验

从表 7 可以看出，当特征模板的窗口半径设

[1] http://www.chokkan.org/software/crfsuite/

[2] 形态切分中得到的第 1 个词素为该单词的词干。

[3] http://morpho.aalto.fi/projects/morpho/

[4] http://www.nltk.org/_modules/nltk/tag/hmm.html

[5] http://homepages.inf.ed.ac.uk/lzhang10/maxent.html

置为 2 时，实验结果得分最低。随着窗口半径的增大，实验结果也明显提高，当窗口半径设置为5时，实验结果得分最高。当窗口半径超过 5 以后，实验结果几乎持平，并带有轻微的下降。通过对实验数据进行分析，对该现象的解释如下。

1) 过拟合：特征模板的窗口半径超过 5 以后，系统对未登录词的预测能力开始下降，说明系统中存在过拟合现象，只能有效预测训练集中出现的字符序列，而对训练集中没有出现的字符序列的处理能力下降。

2) 兼类词歧义：当兼类词构形时，词缀作为上下文特征，判断其词性时发挥重要作用。如果兼类词没有发生构形，此时没有词缀可以作为其特征信息，因此系统把该词概率最大的词性作为它的词性返回，如单词“باشقا”(其他，下次)在训练集中出现 69 次，其中作为形容词出现了 37 次，作为副词出现了 7 次，作为语气词出现 25 次，因此在测试集中作为副词出现时，模型错误的预测成形容词。

3) 切分歧义：当不同的词干构形之后得到相同的单词时，把出现频率最高的切分形式作为预测值。如副动词“بېرىپ”的切分形式有“بار+ىپ”(去)和“بەر+ىپ”(给)，其第一种切分形式在训练集中出现 9 次，第二种切分形式出现 28 次，因此把测试集中出现的所有“بېرىپ”切分成第二种形式。

4) 音变还原歧义：构形过程中发生音变现象的词缀还原时，还原成同一类词缀的另一种形式。如测试集中单词“ئىشلىرىدىمۇ”(在事业当中)的正确切分形式为“ئىش+لار+ى+دا+مۇ”，但测试结果中的切分形式为“ئىش+لار+ى+ده+مۇ”，没能正确还原时位格缀“دا”上发生的弱化现象，虽然在训练集中时位格缀“دا”和“ده”发生弱化次数同样为 3 次，但是字符“ا”发生 267 次弱化，其弱化次数明显小于字符“ە”弱化的次数(313)，因此模型没能正确还原这种音变现象。

针对第一种情况，需要合理设置特种模板的窗口半径，而针对其他 3 种情况，通过单词间的上下文关系，可以缓解歧义现象。本文只考虑词内字符间的上下文关系，如果把单词间的上下文关系作为特征信息参与训练和测试，可以降低歧义造成的误判率。

3.3.2 对比实验

从 4 组实验结果(表 8)中可以发现，实验 1 的结果最好，实验 2 和 3 的结果明显低于实验 1，实验 4 的结果接近于实验 1。分析其原因如下。

1) 实验 2 中的 Morfessor 模型训练时需要大量的语料，而我们的训练语料的规模不大。实验 2 的结果中发现过度切分现象，如词干“كەسىپ”(职业)被错误地切分成“كەس+ىپ”的形式，是因为在语料库中出现由词根“كەس”(切)构词的不同单词。由于 Morfessor 模型不具备标注能力，因此实验中没有形态标注和音变还原结果。

2) 实验 3 中词素切分和词干提取的结果不理想，是因为 HMM 模型假设当前状态只跟前一个状态有关，因此没有充分利用字符序列中的上下文信息。

3) 实验 4 中 MEM 模型采用局部最优化，而实验 1 中的 CRF 模型采用全局最优的训练模式，因此 MEM 对训练集中未出现情况的处理能力比 CRF 模型差。图 4 给出当观察窗口半径取值不同时，CRF 和 MEM 模型未能处理的未登录词数量的区别。

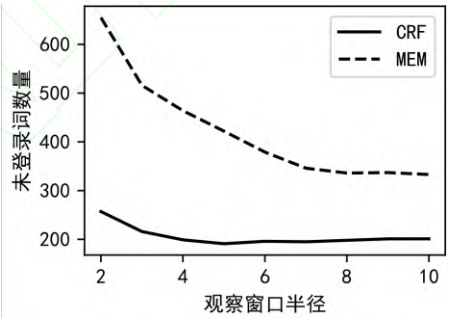

图 4　未登录词数

Fig. 4　Count of out of vocabulary words

4 结论

本文提出了基于字符级别的有监督维吾尔语形态协同分析方法。该方法结合维吾尔语的黏着性特点，把词素边界识别，形态标注及音变还原等形态分析任务定义为字符序列的标注问题，采用序列标注方法实现了用一个模型完成复杂形态分析的任务。实验结果证明，我们提出的模型在维吾尔语的形态分析任务中得到了较好的效果，并且根据不同的应用场景，从结果中可以得到词干、词性标注等不同的分析数据。该模型在相似语种之间具有一定的通用性，因此还可以用于形态特征跟维吾尔语相似的哈萨克语、柯尔克孜语等语种的形态分析任务中。针对实验过程中发现的问题。在后续工作中，将利用单词间的上下文

关系作为特征进行模型优化，从而能有效降低歧义导致的误判率，进一步提高形态分析的正确率。

参考文献

[1] 宗成庆. 统计自然语言处理. 2 版. 北京: 清华大学出版社, 2013: 8–9

[2] Harris Z S. From phoneme to morpheme. Language, 1955, 31(2): 32–67

[3] Harris Z S. Morpheme boundaries within words: report on a computer test. papers in structural and transformational linguistics. Dordrecht: Springer, 1970: 68–77

[4] Merialdo B. Tagging English text with a probabilistic model. Computational Linguistics, 1994, 20(2): 155–171

[5] Goldsmith J. Unsupervised learning of the morphology of a natural language. Computational Linguistics, 2001, 27(2): 153–198

[6] Creutz M, Lagus K. Lindén K, et al. Morfessor and hutmegs: unsupervised morpheme segmentation for highly-inflecting and compounding languages // Second Baltic Conference on Human Language Technologies. Tallinn, 2005: 107–112

[7] Creutz M, Lagus K. Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing, 2007, 4(1): 3:1–3:34

[8] Ruokolainen T, Kohonen O, Virpioja S, et al. Supervised morphological segmentation in a low-resource learning setting using conditional random fields // Conference on Computational Natural Language Learning. Sofia, 2013: 29–37

[9] 玉素甫·艾白都拉, 吾守尔·斯拉木,赛依提·阿不都拉. 维语词法分析器研究成功. 中文信息, 1997(4): 31–35

[10] 阿依克孜·卡德尔, 开沙尔·卡德尔, 吐尔根·依布拉音. 面向自然语言信息处理的维吾尔语名词形态分析研究. 中文信息学报, 2006(3): 43–48

[11] 艾山·吾买尔. 维吾尔语名词词干提取算法的研究. 中国中文信息学会信息检索与内容安全专业委员会 // 第 4 届全国信息检索与内容安全学术会议论文集(上). 北京, 2008: 180–186

[12] Aishan W, Zaokere K, Parida T, et al. Maximum entropy combined FSM stemming method for Uyghur // 2009 Oriental COCOSDA International Conference on Speech Database and Assessments. Urumqi, 2009: 51–55

[13] 薛化建, 董兴华, 王磊, 等. 基于词缀库的非监督维吾尔语词切分方法. 计算机工程与设计, 2011, 32(9): 3191–3194

[14] 麦热哈巴·艾力, 姜文斌, 王志洋, 等. 维吾尔语词法分析的有向图模型. 软件学报, 2012, 23(12): 3115–3129

[15] 张海波, 蔡洽吾, 姜文斌, 等. 基于联合音变还原和形态切分的形态分析方法. 中文信息学报, 2014, 28(6): 9–17

[16] 米尔阿迪力江·麦麦提. 基于 Morfessor 的维吾尔语词干提取和词性标注的研究. 乌鲁木齐: 新疆大学, 2015

[17] Tursun E, Ganguly D, Osman T, et al. A semisupervised tag-transition-based Markovian model for Uyghur morphology analysis. ACM Transactions on Asian and Low Resource Language Information Processing, 2016, 16(2): 8:1–8:23

[18] 哈里旦木·阿布都克里木, 程勇, 刘洋, 等. 基于双向门限递归单元神经网络的维吾尔语形态切分. 清华大学学报(自然科学版), 2017, 57(1): 1–6

[19] Maimaiti M, Wumaier A, Abiderexiti K, et al. Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. Information-an International Interdisciplinary Journal, 2017, 8(4): 157

[20] 米莉万·雪合来提, 刘凯, 吐尔根·依布拉音. 基于维吾尔语词干词缀粒度的汉维机器翻译. 中文信息学报, 2015, 29(3): 201–206

[21] Mi C, Yang Y, Dong R, et al. Optimized Uyghur segmentation for statistical machine translation // International Conference on Applications of Natural Language to Information Systems. Cham, 2015: 395–398

[22] 吐尔洪·吾司曼, 维尼拉·木沙江. 维、哈、柯多语种搜索引擎中索引器的研究. 新疆大学学报(自然科学版), 2011, 28(2): 132–135

[23] Lafferty J D, Mccallum A, Pereira F, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data // International Conference on Machine Learning. Williamstown, 2001: 282–289

[24] 李航. 统计学习方法. 北京: 清华大学出版社, 2012: 194–208