文章编号：1003-0077(2014)06-0009-09

基于联合音变还原和形态切分的形态分析方法

张海波，蔡洽吾，姜文斌，吕雅娟，刘 群

（中国科学院计算技术研究所 智能信息重点实验室，北京 100190）

摘 要：传统的形态分析方法，一般是先进行音变还原工作，再进行形态切分工作。音变还原工作的好坏直接影响形态切分工作的优劣，两者之间存在错误传播的问题。鉴于传统形态分析方法存在的错误传播问题，该文提出了基于联合音变还原和形态切分的形态分析方法。该方法通过使用具有双重功能的联合标签，同时实现了音变还原及形态切分的功能。由于该方法不依赖于黏着语的特有的语言学规则，因此便于扩展到新的语言上。结果表明，联合音变还原和形态切分的形态分析方法要优于传统的先进行音变还原后形态切分的形态分析方法，能够很好地解决先音变还原后形态切分带来的错误传播问题。

关键词：形态分析，音变还原，形态切分

中图分类号：TP391 **文献标识码**：A

A Joint Voice Harmony Restoration and Morphological Segmentation Model for Morphological Analysis

ZHANG Haibo, CAI Qiawu, JIANG Wenbin, LV Yajuan, LIU Qun

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In order to solve the problem of error propagation in traditional morphological analysis method with a pipline of the voice harmony restoration and the morphological segmentation, this paper presents a unified approach combining voice harmony restoration and morphological segmentation. It makes use of a kind of integrated label for both the voice harmony restoration and morphological segmentation. Experiments show that the proposed method can improve precision and alleviate the error propagation in traditional morphological analysis method.

Key words: morphological analysis, voice harmony restoration, morphological segmentation

1 引言

黏着语是一种通过在词干基础上粘贴不同的词缀来实现语法功能的语言类型。我国的很多少数民族语言，如维吾尔语、蒙古语、哈萨克语、朝鲜语等都属于黏着语。黏着语在我国分布广泛且使用人口众多，分布地区具有较高的政治经济价值。黏着语每个词的变化形式最多可达数百种，甚至上千种。现有主流的机器翻译方法基本上不考虑词形变化，把每个不同词形的词都当成独立的词语来考虑。但是对于黏着语而言，这种做法就会带来比较严重的问题。因为这类语言的词语变化非常灵活，形式多样，这样会导致机器翻译时出现大量未登录词，严重影响机器翻译的性能。

黏着语形态分析一般包含音变还原和形态切分两个子任务。音变还原是黏着语形态分析中的重要基础处理环节。音变现象是词干与词缀连接时发生弱化、增音、脱落等现象。据统计（表 1），可以看出，音变现象在 3 种语言中广泛存在。音变现象使形态分析工作变得更加困难，音变问题解决的好坏直接影响后续的形态分析工作。音变还原现象研究界关注较少，艾山·吾买尔等人[1]提出基于噪声信道的识别模型，该模型以弱化的词干词尾的二个字符、3 个字符以及最后音节作为上下文，建立相应的语言模型以及似然度计算公式来解决音变现象问题。麦

收稿日期：2014-06-20 **定稿日期**：2014-08-20

热哈巴·艾力等人[2]提出了基于最大熵模型的音变还原模型，该模型主要基于词性标注工作的思想，首先找出每一个维吾尔语字母可能对应的集合，然后通过维特比算法对维吾尔语字母进行线性序列标注，这种方法避开了总结和制定音变现象复杂的规则。然而上述方法只考虑了当前字母的上下文环境，未考虑形态切分任务对于音变还原的影响，并且音变还原很难做到百分之百的准确率，存在错误传播的问题。

表1 音变现象统计

语 种	句子数	词 信 息		
		# token	# diff	比例/%
维吾尔语	71 287	1 099 319	142 621	12.97
蒙语	13 965	236 084	12 739	5.39
韩语	53 272	715 808	107 663	15.04

形态切分是形态分析任务中重要的环节，主要对黏着语单词的各种切分方式进行排歧，选择最优的切分方式。当前很多形态切分都是以序列标注模型求解的，典型的工作有赵伟[3]等提出的运用条件随机场解决蒙古语中的词语切分问题，该工作将蒙古语词内的每一个字母看成最小的切分单元，对单词内的每一个字母进行BMES标注，B代表Begin，M代表Middle，E代表End，S代表Single，进行标注的时候只提取当前字母左右的几个字母窗口范围内的特征，然后运用条件随机场进行判别式训练，然后采用维特比算法进行求解出最优的标注序列，进而得到蒙古语词的切分结果。然而上述方法只考虑了当前待标注字母上下文的环境，并未考虑音变还原的相关信息，并且形态切分的输入是音变还原之后的结果，音变还原性能的优劣直接影响形态切分的效果。

为了解决音变还原任务中未考虑形态切分的信息，形态切分任务中未考虑音变还原的信息，以及音变还原和形态切分任务之间存在错误传播的问题。

黏着语的形态分析工作较为复杂，其困难主要表现在如下3点。

1. 音变现象严重

音变现象是词干与词缀连接时发生弱化、增音、脱落等现象。以维吾尔语为例，说明各种音变现象[4]。

(1) 弱化现象

弱化现象是词干接词缀时词干中的某些字母会转换成其他字母的现象。弱化现象不仅出现在元音上，也会出现在辅音上。当词缀层次多时，弱化现象也会出现在词缀所包含的字母上。例如，

mektep(学校，词干)+im(第一人称单数，词缀)=mektipim(我的学校)

其中词干中的第二个元音e弱化为i。

(2) 增音现象

增音现象是词干接词缀时，会增加一个字母的现象。

(3) 脱落现象

脱落现象是词干接词缀时有些字母会出现脱落的现象。

本文提出了联合音变还原与形态切分的模型进行解决，该模型以序列标注为基本框架，自动地通过对齐发掘还原以及切分规律，在抽取的实例上训练感知机分类器，将音变还原以及形态切分融合在一个任务中完成，在复杂度变化不大的情况下，显著地提高了形态分析的质量。

实验结果表明，在维吾尔语、蒙古语以及韩语上显示，联合模型大幅度领先于传统的先进行音变还原后形态切分的模型以及有向图模型[4]。文本组织结构安排如下，第2部分阐述什么是黏着语的形态分析，第3部分阐述相关工作，第4部分阐述基于字符分类的音变还原，第5部分阐述基于字符分类的形态切分，第6部分阐述联合音变还原及形态切分的模型，第7部分阐述相关实验，最后对文章进行总结。

2 黏着语形态分析

黏着语是一种通过在词干基础上粘贴不同的词缀来实现语法功能的语言类型。对于黏着语而言，由于词语变化非常灵活，形式多样，通过在词干的基础之上不断添加词缀来表达语法意义，词的构成方式如图1所示。因此，黏着语形态分析的侧重点在于词干与词缀的切分工作。以维吾尔语为例，说明黏着语的形态分析主要任务。如图2所示，维吾尔语形态分析的输入是维吾尔语单词组成的句子，经过形态分析之后，输出词干与词缀切分之后的维吾尔语句子。

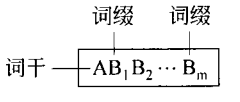

图1 黏着语的单词构成方式

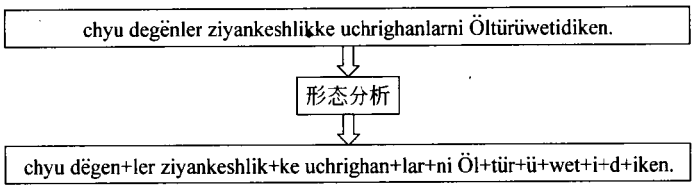

图 2　维吾尔语形态分析示意图

(4) 组合情况

组合情况是以上几种情况会同时出现。

黏着语的音变现象较为严重，据统计，韩语中发生音变的词占总词数的 15%左右，维吾尔语中发生音变的词占总词数的 13%左右。因此，黏着语的形态分析工作面临的首要任务就是如何将发生音变现象的词进行还原，该问题统称为音变还原问题。

2. 词干词缀的切分歧义

黏着语的形态分析面临词干词缀切分带来的歧义，主要表现于如下两个方面。

(1) 同一个词提取的词干具有歧义[4]

(2) 同一个词提取的词缀具有歧义

3. 语言资源贫乏

黏着语的形态分析任务面临的第 3 点困难是缺乏语言资源，没有大规模的形态分析标注语料。在仅有的小规模标注的形态分析语料上，并且没有统一的标注标准，各单位、各组织根据自己制定的标准进行标注，造成资源的进一步匮乏。如何在小规模语料资源上进行黏着语的形态分析任务是一项具有挑战的任务。

3　相关工作

根据知识表示和知识获取方式的不同，可以将形态分析的方法分成基于规则和基于统计两类。根据学习方法的不同，后者还可以进一步分为参数方法(或统计方法)等。

1. 基于规则的方法

规则的方法主要是基于维吾尔语的特有语言学规则进行音变还原、词干词缀切分等的方法，一般由语言学家根据每一种语言特点，制定相应的音变还原规则、词干词缀切分规则或者提取相应的词典，然后应用于形态分析任务中。

古丽拉·阿东别克[5]提出了以“词＝词根＋附加成分”的结构，对维文词的词法和语法结构进行了归纳，提出了维吾尔语词切分的一些规律和实现方法。维吾尔语相关语言学专家对维吾尔语名词、动词等词类的形态变化规律进行总结[6]，艾山·吾买尔[7]和早克热·卡德尔[8]等人借鉴了这些已有的总结规律。

上述基于规则的形态分析方法，存在如下缺点：第一，必须依靠语言学家制定相关的语言学规则，耗费了大量的人工成本，并且时间周期较长。第二，随着规则的不断增多，规则之间难免会发生顾此失彼的冲突现象，规则描述的粒度也会越来越细，越来越不容易维护和管理。第三，规则难以解决存在的一些歧义现象，比如词干切分引起的歧义。由于上述缺点，基于规则方法的形态分析器性能比较低，而且难以维护。

2. 基于统计模型的方法

基于统计模型的形态分析方法，利用标注好的语料库提取大量的特征或者统计相应的概率，然后运用最大熵模型、CRF 模型、语言模型等进行训练，通过维特比算法进行解码，从中求得一条最优的路径，该路径即为最优的形态分析路径。

对于维吾尔语而言，麦热哈巴·艾力[2]将音变还原问题转化为单词内部字母的标注问题，以单词为单位，训练字母在音变后的字母候选规律及其概率，然后使用最大熵模型进行训练，利用维特比算法进行维吾尔语词的标注，从而得出音变还原之后的词。麦热哈巴·艾力[4]提出了有向图模型运用在维吾尔语形态分析的方法，该方法采用了两个决定的关系，当前的词干仅由上一个词干决定，当前的词缀仅由当前词上一个词缀决定，首先通过已有的词干词缀词典进行枚举词的可能切分候选，然后采用双层的语言模型进行切分的排除歧义工作。

基于统计模型的形态分析方法有以下优势：首先，由于直接从语料库中获取形态切分及音变还原的知识，不需要人工调试规则和词典；其次，由于形态切分及音变还原的知识是直接来源于真实的标注语料库，所以与真实情况的切分尽量保持一致；最后，由于是带参数的机器学习，形态分析与语言本身无关，所以形态分析模型可以迅速迁移到新的语言上。

然而，上述基于统计模型的方法存在如下缺点。在形态切分及排除歧义之前必须要进行相应的音变还原工作，音变还原工作的准确率和速度直接影响下一步形态切分的性能，存在错误不断向下传播的问题，如图3所示。音变还原以及形态切分两个任务没有互相考虑对各自任务的影响。

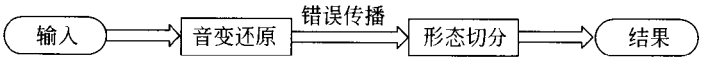

图3 传统形态分析工作错误传播现象

鉴于上述基于统计模型的形态分析方法存在错误传播的问题，音变还原任务中未考虑形态切分的信息，以及形态切分任务中未考虑音变还原的信息，本文提出了联合音变还原和形态切分的方法，能够同时解决音变还原问题以及形态切分问题，不存在错误传播的问题。

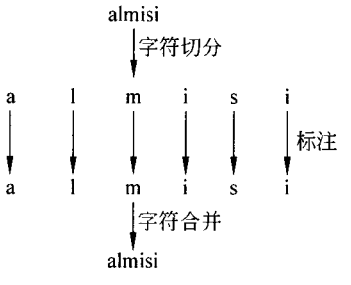

图4 音变还原转变成字符标注问题

4 基于字符分类的音变还原

本部分阐述基于感知机模型的音变还原工作。该模型以序列标注为基本框架，自动地通过对齐发掘还原以及切分规律，在抽取的实例上训练感知机分类器，实现音变还原任务。

4.1 音变还原中的感知机训练和维特比解码算法

传统意义上的感知机常用于解决两类的分类问题，然而，在自然语言处理任务中的分类种类通常大于两类。对于音变还原任务而言，需要判断词内每个字母的所属分类，根据字母的分类情况，产生音变还原结果。我们可以通过词内字母对齐算法得到每一个字母可能的分类标签，关于词内字母对齐算法将在下一小节进行论述。因此音变还原问题转换为自然语言处理任务中常见的词性标注工作。对于单词内的每一个字母进行分类标注，然后组合标注，即可得到音变还原的结果。如图4所示，以维吾尔语单词almisi为例，首先对该单词进行字符切分，变成由字符组成的序列，此时对该序列中的每一个字符进行标注，然后组合标注后的结果即可，该结果就是音变还原的结果为almasi，可以发现i变成了a，发生了音变现象。

为解决传统感知机的上述问题，可以对其进行转化成多元分类问题。对于每个字母，通过模型分别计算这字母在所属所有类别时的模型得分，然后选择最高分的类别作为这个字母的最终类别。序列标注的权重需由感知机模型进行训练得到。每次的解码过程，我们采用维特比算法对基本字母序列进行序列标注。整个训练过程如图5所示。

设输入待分析的句子的原子序列为 $x_i \in X$，输出标注序列 $y_i \in Y$，X 表示训练语料中的所有句子，Y 表示对应的标注。其中 $GEN(x)$ 采用维特比算法得到输入句子 x_i 的候选标注结果，$\Phi(x_i, z)$ 为输入句子的特征向量，然后计算 $\Phi(x_i, z) \cdot w$ 得分，选择最高的标注序列 z。y_i 表示正确的标注序列。

算法1 平均感知机算法

1: 输入：训练实例(x_i, y_i)
2: $w_0 \leftarrow 0$
3: For t $\leftarrow$ 1..T, i $\leftarrow$ 1..N do
4: $z_i \leftarrow \text{argmax}_{z \in GEN(x_i)} \Phi(x_i, z) \cdot w_j$
5: $w_{j+1} \leftarrow w_j + \Phi(x_i, y_i) - \Phi(x_i, z_i)$
6: 输出：$w \leftarrow \sum_j w_j$

图5 平均感知机训练算法

4.2 词内字母对齐算法

在形态丰富的音变还原工作过程中，我们需要提取到每一个字母对应的标注标签，换句话而言就是每一个字母对应的tag的集合。我们需要设计一个算法抽取出每一个字母对应的标签集合。该算法可以参照文献[2]。

4.3 特征设计

感知机训练的特征模板，如表2所示。其中，C_0 表示当前字符，当前字符左边的第一个字符为 C_{-1}，同理，当前字符右边的第一个字符用 C_1 表示。

表 2　音变还原的特征模板

$C_n(n=-5..5)$	$C_nC_{n+1}(n=-5..4)$	$C_nC_{n+1}C_{n-2}(n=-5..3)$
$C_0C_n(n=-5..5)$	$C_0C_nC_{n-1}(n=-5..4)$	$C_nC_{n-1}C_{n+2}C_{n-3}(n=-5..2)$

5　基于字符分类的形态切分

基于字符分类的形态切分，主要是基于线性词语表示方法。基于线性词语表示方法，就是将每一个黏着语的单词看成一个线性结构，句子中整个单词序列也构成一个线性序列。如图 6 所示。

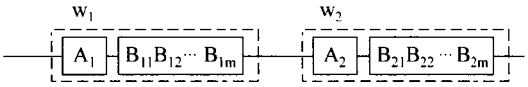

图 6　线性词语表示

从图 6 可以看出，w_1 到 w_2 是一种线性结构，w_1 内部的词干到词缀是一种线性结构。线性结构的优点就是结构简单，容易进行建模以及编程实现。基于线性结构表示方法的典型形态分析工作就是线性序列标注模型。该线性序列标注模型通常将单词内部的每一个字母看成是一个独立的单元，通过提取每一个字母的上下文信息，对其进行 BMES 标注，进而可以对该单词进行词干以及词缀的切分工作，最终得到形态分析的结果。其中，BMES 代表的含义如表 3 所示。

表 3　BMES 含义表

标签	含　义
B	词干或者词缀的开始字母
M	词干或者词缀的中间字母
E	词干或者词缀的结尾字母
S	单个字母组成的词干或者词缀

我们以蒙古语句子为例，说明序列标注模型，如图 7 所示。

在上面这个图中，输入的是蒙古语的句子，首先需要进行原子序列的切分，单词之间加“#”，单词内的每一个字母之间加空格，进而切分成一个线性的词语表示结构，满足单词之间是线性结构，单词内部的字母之间是线性结构。其次，在切分好的线性序列的基础之上，对于每一个原子提取上下文特征信息，利用 BMES 进行标注，可以得到标注的结果，结果是由 BMES 组成的一个序列。最后，通过 BME 或者 BE 或者 S 组成一个单元的方式进行切分蒙古语句子。

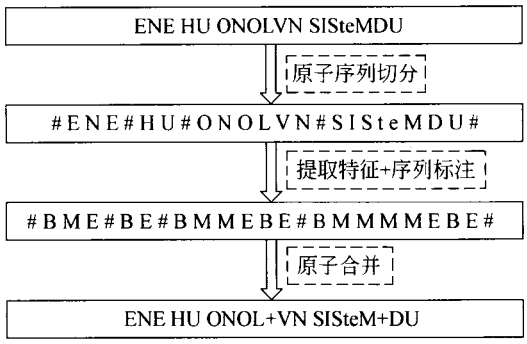

图 7　序列标注模型过程示意图

在本部分我们用感知机模型进行训练，采用表 2 的特征模板，解码算法采用维特比算法。

6　联合音变还原与形态切分

联合音变还原和形态切分方法，将音变还原问题和形态切分问题融合成一个序列标注的问题。一次序列标注既可以得到音变还原的结果，也可以得到形态切分的结果，解决了传统的形态分析方法存在错误传播的问题。本部分首先阐述联合音变还原和形态切分的方法的原理，然后阐述联合方法的训练及解码框架，最后阐述采用的特征模版。

6.1　联合标签的生成

联合音变还原和形态切分的方法，同时进行音变还原工作和形态切分工作，通过一次的序列标注，达到双重效果。该方法采用一种联合标签，该标签具有音变还原的功能，同时也具有形态切分的功能。

以维吾尔语单词“almisi”为例说明。首先，单词“almisi”的原始形式为“almasi”，通过词内字母对齐算法[2]，生成如下对应关系，如图 8 所示。

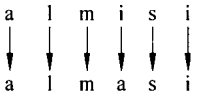

图 8　词内字母对齐

生成标签<a, a>　<l, l>　<m, m>　<i, a>　<s, s>　<i,i>。其次，通过单词的形态分析形式“alma+si”，生成如下标签，<a, B>　<l, M>　<m, M>　<i, E>　<s, B>　<i,E>。

最后，进行组合标签，如图 9 所示，得到<a, B-a>　<l, M-l>　<m, M-m>　<i, E-a>　<s, B-s>　<i,E-i>。

6.2　训练框架

该方法的训练流程如下：首先，对于语料中的

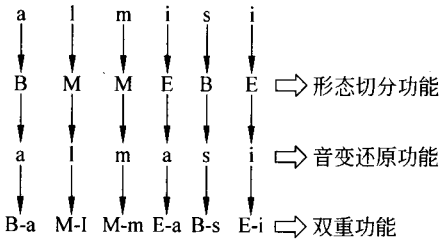

图 9 联合标签生成过程

原始词以及形态分析之后的词，利用词内字母对齐算法进行抽取对齐字母对，对齐字母对目标端即为源端字母的标注标签。其次，根据形态分析之后的词，对于原始词内的每一个字母进行 BMES 标注。然后，将每一个字母的对齐字母对标注以及 BMES 标注进行组合成新的标注标签，该标签是音变还原以及切分标注。BMES 的组合，因此该标签具有双重的功能。最后，提取单词内每一个字母的上下文信息，并生成相应的特征文件，利用感知机进行训练，得到模型文件。训练框架图如图 10 所示。

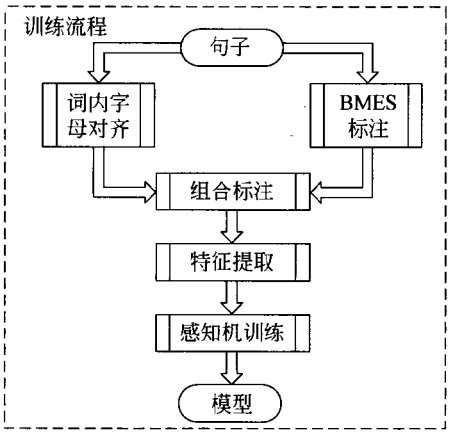

图 10 联合模型的训练框架

6.3 特征设计

特征设计是基于联合音变还原和形态切分方法的关键因素。特征设计的优劣直接影响系统的性能。本文采用的特征模版如图 11 所示，其中，表示当前考察的字母，表示当前考察字母左边第 n 个字母，表示当前考察字母右边第 n 个字母。

类型	特征模板	
1 元组	C_i	$-4 \leqslant i \leqslant 4$
2 元组	$C_{i-1} \circ C_i$	$-3 \leqslant i \leqslant 4$
3 元组	$C_{i-2} \circ C_{i-1} \circ C_i$	$-2 \leqslant i \leqslant 4$
4 元组	$C_{i-3} \circ C_{i-2} \circ C_{i-1} \circ C_i$	$-1 \leqslant i \leqslant 4$

图 11 联合模型系统采用的特征模板

6.4 解码框架

联合模型系统解码框架示意图如图 12 所示。首先，按照相同的前处理，将输入的句子处理成原子序列及“#”组成的序列，然后对该序列进行相应的联合标签标注。解码算法采用维特比算法，得到最优的形态切分的结果。在求得结果中，两个“#”之间的部分是一个单词，根据两个“#”之间的 BMES 标注可以对该单词进行切分，利用对齐对标签可以恢复单词的准确形式，进而得到联合音变还原和形态切分模型的形态分析的结果。

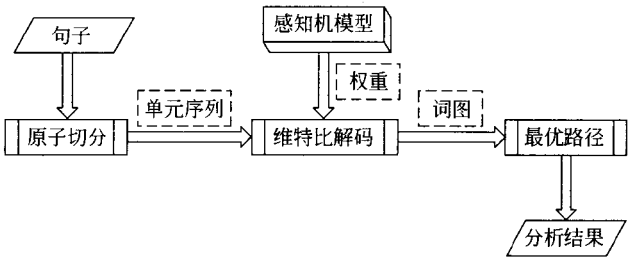

图 12 联合模型系统解码框架

7 实验和分析

7.1 实验数据

本章的维吾尔语的语料资源是由新疆多语种重点实验室手工标注的《维吾尔语百万词词法分析语料库》，此语料库包括 72 741 个完整的句子，分别选择 1%作为开发集和测试集，开发集共包含 727 句，测试集包含 727 句。剩余的 99%作为训练集，共包含 71 290 句。蒙语是由内蒙古大学蒙古学学院开发的 20 万词规模词法分析语料库上进行实验。该语料库共包括 14 115 个完整的句子，我们从中随机抽取出各 5%的语句分别用做开发集和测试集，各含 705 句，剩余 90% 的语句用做训练集，含 12 705 句。韩语语料由网络资源以及与合作方共同开发的资源组成，共包含 54 358 句，分别选择 1%作为开发集和测试集，开发集共包含 543 句，测试集包含 543 句。剩余的 99%做为训练集，共包含 53 272 句。

黏着语的词法分析结果结构远比汉语复杂，传统的正确率、召回率和 F 值不能直接适用。本工作中我们定义和采纳了多种指标，从不同角度和层面考量词法分析器的性能。

a) 词级正确率

以词为单位计量，仅当词内词干、词缀及其标注均正确时，该词才是分析正确的。

$$Pw = \frac{\text{系统输出中正确结果的词数}}{\text{系统输出全部结果的词数}} \times 100\% \quad (1)$$

b) 词干词缀级正确率 Psa，召回率 Rsa 和 Fsa 值

$$Psa = \frac{\text{系统输出中正确结果的数目}}{\text{系统输出全部结果的数目}} \times 100\% \quad (2)$$

$$Rsa = \frac{\text{系统输出中正确结果的数目}}{\text{测试语料的正确答案的数目}} \times 100\% \quad (3)$$

$$Fsa = \frac{2 \times Psa \times Rsa}{Psa + Rsa} \times 100\% \quad (4)$$

以词干和词缀为单位计量，仅当词干或词缀及相应标注正确时，该词干或词缀才是分析正确的。因此，词干和词缀可类比为汉语词法分析中的词。此评价标准引自文献[9]。

7.2　基于感知机的字符分类音变还原

本部分实验我们将在维吾尔语、蒙古语以及韩语 3 个语种上进行实验，实验结果主要通过词级的准确率来进行衡量，词级准确率是还原正确的词数目与总词数目的商得到。实验结果如表 4 所示。

表 4　音变还原的性能

语　种	P_w/%	语　种	P_w/%
维吾尔语	93.14	韩语	95.01
蒙古语	99.56		

从上表中可以看出，在维吾尔语上，系统的准确率达到了 93.14%。对于蒙古语而言，基于平均感知机的音变还原模块的准确率为 99.56%，识别的准确率相当高。对于韩语而言，基于平均感知机的音变还原模块的准确率为 95.01%。

本文传统的音变还原部分采用的方法类似于麦热哈巴[2]的工作，麦热哈巴在 6 万多语料规模上，自动还原模型对测试对象中整个词的还原正确率达到了 90%。本文在 7 万多语料规模上，中缀音变还原模型对测试对象中整个词的还原准确率达到 93.14%。从结果上可以看出，本文提出的方法相对于前人较好的音变还原工作具有可比性，并且能够取得更为有效的性能。

7.3　基于感知机的字符分类的形态切分

此实验的目的是为了测试基于感知机的字符分类的形态切分的性能。该实验的输入是经过准确音变还原之后的标准数据，排除音变的影响，单独测试形态切分的性能。系统的性能如表 5 所示，可以看出，形态切分的性能几乎都能达到 98% 以上，达到了很高的性能。

表 5　形态切分的系统性能

语种	P_w/%	P_{sa}/%	R_{sa}/%	F_{sa}/%
维吾尔语	98.82	98.78	98.74	98.76
蒙古语	98.71	98.77	98.64	98.71
韩语	97.30	97.95	98.00	97.98

7.4　基于联合方法的形态分析的性能要优于传统的形态分析

此实验的目是为了验证联合音变还原和形态切分的形态分析性能要优于传统的形态分析方法。基线系统为先进行音变还原，然后在进行形态切分的系统。基线系统的音变还原将产生的 1best 结果输出给形态切分系统。基线系统设置音变还原的 1best 结果，是通过实验进行验证，随着音变还原的 nbest 结果增多，系统逐渐下降，因此此处设置为 1best。为了排除由于不同的特征模板带来实验效果的差异，基于联合方法采用的特征模板与基线系统采用的特征模板保持一致。测试结果如表 6 所示，可以看出，基于联合方法的系统，在维吾尔语上，词级准确率提升了 0.64 个点，在蒙古语上，词级准确率提升了 0.42 个点，在韩语上，词级准确率提升了 5.89 个点。测试结果表明，基于联合音变还原和形态切分的形态分析方法要比传统的先音变还原后形态切分的方法要好，可以很好地解决错误传播的问题，音变还原和形态切分两个任务互相影响。

表 6　联合模型的系统性能

语　种	系　统	P_w/%	P_{sa}/%	R_{sa}/%	F_{sa}/%	速度/(词/s)
维吾尔语	基线系统	93.60	**94.92**	92.73	93.81	305
	联合系统	**94.24**	94.25	**94.68**	**94.47**	79

续表

语 种	系 统	P_w/%	P_m/%	R_m/%	F_m/%	速度/(词/s)
蒙古语	基线系统	96.10	96.98	95.35	96.16	612
	联合系统	**96.52**	**97.07**	**97.26**	**97.16**	267
韩语	基线系统	88.71	92.86	90.67	91.75	59
	联合系统	**94.60**	**96.43**	**96.18**	**96.30**	22

在韩语上，系统的准确率出现了大幅度上升，产生此种现象的主要原因是在韩语中大约30%以上的词出现了音变现象，高度的音变现象导致了音变还原之后的结果有很多错误，导致基线系统的系能很低。然而，联合模型的系统由于将音变还原和形态切分两个任务联合起来实现，解决了音变还原任务错误传播的问题，大幅度地提升了系统的性能。

对于速度而言，在3个语种上，联合系统都明显低于基线系统。通过研究，我们发现，联合系统由于将音变还原标签和形态切分标签联合起来组成联合标签的原因，导致了存在了大量的分类标签，这些大量的分类标签导致了大量的查询以及增加解码空间，从而导致速度明显低于基线系统。

8 总结

本文提出了一种联合音变还原和形态切分的形态分析方法。该方法通过使用具有双重功能的联合标签，同时实现了音变还原及形态切分的功能。由于该方法不依赖于黏着语的特有的语言学规则，因此便于扩展到新的语言上。实验结果表明，基于联合音变还原和形态切分的形态分析方法要优于传统的先进行音变还原后形态切分的形态分析方法，能够很好地解决先音变还原后形态切分带来的错误传播问题。

参考文献

[1] 艾山·吾买尔，吐尔根·依布拉音. 基于噪声信道模型的维吾尔语央音原音识别[J]. 中国计算机语言学研究前沿发展，2010，46(15)：118-120.

[2] 麦热哈巴·艾力，姜文斌，吐尔根·依布拉音. 维吾尔语词法中音变现象的自动还原模型[J]. 中文信息学报，2012，26(1)：91-96.

[3] 赵伟，侯宏旭，从伟，宋美娜. 基于条件随机场的蒙古语词切分研究[J]. 中文信息学报，2010，24(5)：31-35.

[4] 麦热哈巴·艾力，姜文斌，王志洋，吐尔根·依布拉音，刘群. 基于有向图模型的维吾尔语词法分析[J]. 软件学报，2012，23(12)：3115-3129.

[5] 古丽拉·阿东别克，米吉提·阿不力米提. 维吾尔语词切分方法初探[J]. 中文信息学报 2004，18(6)：61-65.

[6] 阿依克孜·卡德尔，开沙尔·卡德尔，吐尔根·依布拉音. 面向自然语言处理信息处理的维吾尔语名词形态分析研究[J]. 中文信息学报，2006，20(3)：43-48.

[7] 艾山·吾买尔，吐尔根·依布拉音，早克热·卡德尔. 维吾尔语名词干提取算法的研究[C]//第四届全国信息检索与内容安全学术会议，2008.

[8] 早克热·卡德尔，吐尔根·依布拉音. 维吾尔语形容词构形词缀有限状态机[J]. 电脑知识与技术，2009，5(4)：937-941.

[9] 宋彦，蔡东风，张桂平，赵海. 一种基于字词联合解码的中文分词方法[J]. 软件学报，2009，20(9)：2366-2375.

[10] 米海涛，熊德意，刘群. 中文词法分析与句法分析融合策略研究[J]. 中文信息学报，2008，22(2)：10-17.

[11] 那顺乌日图，雪艳，叶嘉明. 现代蒙古语料库加工技术的新进展—新一代蒙古语词语自动切分与标注系统[C]. 第十届全国少数民族语言文字信息处理学术研讨会，2005.

[12] 那顺乌日图，淑琴. 面向信息处理的蒙古语规范化探究[J]. 中央民族大学学报(哲学社会科学版)，2006，6：115-122.

[13] 丛伟. 基于层叠隐马尔科夫模型的蒙古语词切分系统的研究[D]. 内蒙古大学硕士毕业论文，2009.

[14] 艳红，王斯日古楞. 基于HMM的蒙古文自动词性标注研究[J]. 内蒙古师范大学报(自然科学汉文版)，2010，39(2)：206-209.

[15] 李文，李淼，张建. 一种带权值参数的非监督式形态分析方法[C]. 少数民族青年自然语言处理技术研究与进展，2010.

[16] 米热古丽·艾力，米吉提·阿不力米提，艾斯卡尔·艾木都拉. 基于词法分析的维吾尔语元音弱化算法研究[J]，中文信息学报，2008，22(4)：43-47.

[17] 米吉提·阿不力米提，等. 维吾尔语中的语音和谐规律及算法的实现[C]. 中国科协2005年会. 2005.

[18] 姜文斌，吴金星，长青，那顺乌日图，刘群，赵理莉，蒙古语词法分析的有向图模型[J]. 中文信息学报，2011，25(5)：94-100.

[19] 阿孜古丽·夏力甫. 维吾尔语动词附加语素的复杂特征研究[J]. 中文信息学报，2008，22(3)：105-109.

[20] 侯宏旭，刘群，那顺乌日图，牧仁高娃，李锦涛. 基于统计语言模型的蒙古文词切分[J]. 模式识别与人工智能，2009，22(1)：108-112.

张海波(1989—)，硕士，主要研究领域为词法分析以及机器翻译。
E-mail：newchance@126.com

姜文斌(1984—)，助理研究员，博士，主要研究领域为词法分析以及句法分析。
E-mail：jiangwenbin@ict.ac.cn

蔡洽吾(1988—)，硕士，主要研究领域为词法分析以及机器翻译。
E-mail：caiqiawu@ict.ac.cn

(上接第 8 页)

[15] Li Z, Zhou G. Unified dependency parsing of Chinese morphological and syntactic structures[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1445-1454.

[16] Zhao H, Huang C N, Li M, et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling[C]//Proceedings of PACLIC. 2006, 20: 87-94.

[17] McDonald R, Crammer K, Pereira F. Online large-margin training of dependency parsers[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 91-98.

[18] Sun W. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1385-1394.

[19] Zhang M, Zhang Y, Che W, et al. Character-Level Chinese Dependency Parsing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 1326 1336.

郭振(1988—)，硕士研究生，主要研究领域为依存句法分析和中文分词。
E-mail：12120416@bjtu.edu.cn

苏晨(1989—)，硕士研究生，主要研究领域为机器翻译。
E-mail：12120447@bjtu.edu.cn

张玉洁(1961—)，教授，主要研究领域为自然语言处理和机器翻译。
E-mail：yjzhang@bjtu.edu.cn