软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(12):3115−3129 [doi: 10.3724/SP.J.1001.2012.04205]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

维吾尔语词法分析的有向图模型*

麦热哈巴·艾力[1,2+], 姜文斌[2,3], 王志洋[2,3], 吐尔根·依布拉音[1], 刘 群[2]

[1](新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046)

[2](中国科学院 计算技术研究所,北京 100190)

[3](中国科学院 研究生院,北京 100049)

Directed Graph Model of Uyghur Morphological Analysis

Mairehaba · AILI[1,2+], JIANG Wen-Bin[2,3], WANG Zhi-Yang[2,3], Tuergen · YIBULAYIN[1], LIU Qun[2]

[1](College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

[2](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100190, China)

[3](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: marhaba@xju.edu.cn

Aili M, Jiang WB, Wang ZY, Yibulayin T, Liu Q. Directed graph model of Uyghur morphological analysis. *Journal of Software*, 2012,23(12):3115−3129 (in Chinese). http://www.jos.org.cn/1000-9825/4205.htm

Abstract: Uyghur is a typical agglutinative language. It has a strong derivational ability with very a rich morphological structure and follows a harmonious rule. In the formation process, some phenomena may occur such as weakened, increased tone and fallen tone. The specific character of Uyghur language determines the difficulty of the Uyghur morphological analysis, including stemming and restoring the changed letter and POS tagging. This paper employs the hierarchical structure of Uyghur word, and proposes a directed graph model for Uyghur morphological analysis. In this model, words and tags are described as a directed graph. In this graph, nodes represent stems, affixes and their corresponding tags, while edges represent the transition, or general probabilities between nodes. Aimed at providing some light on the phenomenon of morphological sandhi in Uyghur language, this paper also proposes a restore model by changing the word to its original form. With the assumption that one letter can be changed to any letter, this model converts restoring problem into a sequence labeling problem, which could be solved by statistical methods. Experiment results on "Mega-words Corpus of Morphological Analysis of Uyghur", which is manually annotated by Xinjiang multilingual key laboratory shows that the accuracy of stemming reaches 94.7%, and the *F* score of stem and affix in line with tag reaches 92.6%.

Key words: Uyghur language; morphological analysis; word segmentation; POS tagging; directed graph

摘 要: 维吾尔语是典型的黏着性语言,其派生能力很强,具有丰富的形态变化,同时遵循语音和谐规律,生成过程中会出现弱化、增音、脱落等音变现象.这些特性决定了维吾尔语词法分析的难点,包括词干提取、发生音变字母的还原以及标注.将维吾尔语词的层次结构引入到词法分析研究中,提出了维吾尔语词法分析的有向图模型,该模型

* 基金项目: 国家自然科学基金(61063026); 国家社会科学基金(10AYY006); 国家工信部电子发展基金(工信部财(2009)553); 新疆高校青年教师科研培养基金(XJEDU2010S07); 新疆大学优秀博士创新项目基金

收稿时间: 2011-04-08; 定稿时间: 2012-02-22

将维吾尔语词法分析描述为有向图结构,图中节点表示词干、词缀及其相应标注,其边表示节点之间的转移或生成概率并将此概率作为候选择优的依据.针对维吾尔语在形态变化过程中发生的音变现象,又提出基于词内字母对齐算法的自动还原模型,该模型将音变现象泛化到每个字母上的假设之下,将还原问题转变成类似于词性标注问题,再利用统计方法进行还原.在对新疆多语种信息技术重点实验室手工标注的《维吾尔语百万词词法分析语料库》上进行的实验中,取得了词干提取正确率为 94.7%,词干与各词缀切分并标注的 F 值达到 92.6% 的好成绩.

关键词: 维吾尔语;词法分析;词语切分;词性标注;有向图

中图法分类号: TP391 **文献标识码**: A

词法分析是自然语言处理的基础,词法分析的优劣直接影响自然语言处理的后续任务,如信息检索、词典编纂、机器翻译等.然而,不同的语言对词法分析有不同的侧重点.对于英语来讲,词法分析的重点是标注,因为英语词具有自然的分隔符——空格,且英语的词缀不多,其形态变化也不复杂,所以英语的形态分析任务不重;汉语词法分析中首先解决的问题是分词,可以说,分词是汉语词法分析的重中之重.

到目前为止,词法分析的研究方法大致可分为基于规则的、基于统计的以及基于混合策略的.基于规则的方法主要依赖语言自身的构词特点与规则,有直观、易实现等优点,同时又有规则不完整、冲突等缺点,且对未登录词的识别没有更好的解决方法,所以其效率并不高;基于统计的方法利用数学模型进行词法分析,虽然需要手工标注一定规模的语料作为训练数据,但结果往往比基于规则的方法好,只是也会出现无法预料的、与实际不符的结果.基于混合的策略则通过引入一些语言规则来约束统计的结果,提高系统的准确率.近年来,研究者提出多种数学模型进行自动词法分析,如:基于隐马尔可夫模型(HMM)的词法分析方法[1]、基于最大熵模型(MEM)[2]以及基于条件随机场模型(CRF)[3]等.这些模型成功地应用到英语、汉语等语言的词法分析研究中,并取得了较好的成绩.

维吾尔语属于阿尔泰语系突厥语族,是典型的黏着性语言.维吾尔语的词干**后可接不同的构形词缀,且可以多层缀接,表示不同的语法功能,同时显现出丰富而复杂的形态变化,如 alma(苹果)、almisi((某人)的苹果)、almining(苹果的)、almiliridin(从(某人的)多个苹果当中)等等.很显然,维吾尔语词法分析的任务之一就是词干和词缀分离,以避免出现诸多未登录词;再者,由于维吾尔语的语音和谐规律,词干后接词缀时有些元音、辅音会弱化成另外一个音(称作弱化)或者出现丢失(称作脱落)、增加(称作增音)等情况.为统一起见,本文中将这些统称为音变现象.因此,为了得到正确的词干与词缀,还需对发生音变现象的字母进行还原;最后,才是词性标注.

维吾尔语的词法分析研究起步较晚,多数研究者采用规则与统计相结合的方法.文献[4]提出利用规则的方法对维吾尔语词进行切分,创建词干库、词缀库,利用向前匹配和向后匹配算法将切分内容与词干库、词缀库进行比较来提取词干和词缀.此方法虽然可行,但会受到词干库、词缀库覆盖率的约束.文献[5−7]提出利用有限状态自动机(FSM)对名词、形容词等静词(非动词)类进行词干提取,其中,文献[5]采用 FSM 与最大熵模型相结合的方法提取名词的词干,其正确率达到 91.2%、回收率为 89.7%、F 值达到 90.51%.这些方法虽对静词(非动词)的词干提取有一定的效果,但是对动词处理就不那么理想.因为维吾尔语是以动词为特征的语言,其动词的语法格式最丰富且最复杂;同时,这些工作把主要精力放到词干提取上,还未考虑词缀.文献[8]提出利用规则对音变现象进行还原,主要是构造词干库、音节库、复合词缀库、单词词缀库,再结合元音弱化的规则作为还原的依据,其实验成功率达到 90%.文献[9]提出通过噪声信道模型的统计方法对音变字母进行还原,实验结果达到了 82.4%.文献[10]通过构建词干库、单词缀库、复合词缀库,采用向前匹配和向后匹配的算法研究了词干(根)与词缀之间的分离,其中,词干(根)与词缀边界的识别率达到了 96%,词干(根)与词缀的分离达到了 92%.文献[11,12]研究了统计模型 CRF 对维吾尔语形态分析的作用,其形态分析正确率达到 92.5%.

总结以上方法,可以发现以下几点:

** 维吾尔语中,词根是指不可再分的最小语义单元;词干是在词根后接构词附加语素后形成的新词;构词词缀是指缀接在词根或词干后,构成新词的附加语素;构形词缀是指缀接在词根或词干后,表示词干或词根新的语法意义的附加语素.

(1) 大多数研究者采用构造词干库、词缀库等词典,从而受到其覆盖面的限制;

(2) 动词与静词的形态分析分开进行,虽然静词的形态分析获得了较好的实验结果,但动词的分析仍未能解决好;

(3) 采用统计方法时,特征的选择主要以字母为主.

针对上述问题,本文提出了维吾尔语词法的有向图模型.此模型中每个词被描述为一个树状结构,根节点表示词干,孩子节点表示词缀,边表示各节点之间的约束关系;一个词串就是各树连接起来的树型结构;词干与词缀所对应的标注也被看成是树状结构,那么整个词串的标注就构成另一个树;利用词干、词缀与对应标注之间的映射关系将两棵树结合起来构成一个平行有向图.通过将词干到词干、词干到词缀、词缀到词缀的(其标注也类似)生成或转移概率作为边的强度来体现各节点之间互相制约关系.另外,针对维吾尔语语音变化的现象,本文也提出了基于词内字母对齐算法的还原模型.

本文第 1 节介绍维吾尔语词法的特点.第 2 节介绍有向图模型.第 3 节介绍音变还原算法的实现过程.最后是实验和结论.

1 维吾尔语词法的特点与难点

维吾尔语的构词、构形都是通过在词根(干)之后接不同的构词、构形词缀来实现,如:

oqu(读,是词根(词干))

oqu+ghuqi(名词构词词缀)→oqughuqi(学生) (构词词缀)

oqu+di(过去式,第三人称(单、复数))→oqudi(他(们)读了) (构形词缀)

oqu+yal(能动式)+ma(否定式)+ywat(现在进行体)+idu(将来时第三人称词缀(单、复数))→oquyalmaywatidu(他(们)没能读出来) (构形词缀)

维吾尔语中构形词缀(由于词法分析以词作为研究对象,所以本文所提及的词缀是指构形词缀)数目很多,其中,名词词缀和动词词缀最多,分别为 49 和将近 200 个[28].词缀的连接可以是多层的,并表现出不同的形态、不同的语法意义.如,以下为动词 al(拿)的几个常用形态:

Al	拿(词干)
Aldim	我拿了
Alalidi	(某人)拿上了
Aldurghanidim	我让别人拿过
Aldurghanliqtin	由于让(他)拿了
…	…

丰富的形态说明维吾尔语词法分析对词干与词缀切分的必要性.

此外,维吾尔语文字属于表音型文字,遵循语音和谐规律,因此在词干、词缀连接时,会发生弱化、增音、脱落等音变现象,见表 1.

Table 1 Phenomenon of morphological sandhi in Uyghur

表 1 维吾尔语音的音变现象

音变现象	例子	说明
弱化	mekt**e**p(学校)+im(第一人称单数)=mekt**i**pim(我的学校) beri**p**(去)+aptu(助动词)=beri**w**aptu((他)已经去了) bal**a**(孩子)+l**a**r(复数词缀)+i(第三人称单数)=bal**ili**ri(他的孩子们)	不仅有元音弱化,还有辅音弱化
增音	arzu(愿望,词干)+um(第一人称单数,词缀)=arzu**y**um(我的愿望)	词干后增加了一个字母 y
脱落	bur**u**n(鼻子,词干)+i(第三人称单数,词缀)=burni(他的鼻子) kong**ü**l(心,词干)+i(第三人称单数,词缀)=kongli(他的心) **a**l(拿)+**i**p+tu+dek=aptudek(好像他拿了)	有时出现多个字母同时脱落的情况
多种现象同时出现	q**a**l (留)+**i**p+t**u**+iken(系助动词)=q**ë**ptiken(听说他留下了)	词干中的 a 弱化成 ë,而 l,i,u 被脱落

语音和谐规律是维吾尔语的一大特点,虽然其他一些黏着性语言,如土耳其语等也存在语音和谐规律,但远

不如维吾尔语复杂.维吾尔语中进行词干、词缀切分时还需要对已发生变化的字母进行还原.

维吾尔语词的形态变化以及音变现象给词干与词缀的切分带来很多不便,总结起来可以概括到以下几类:

(1) 对同一个词进行词干、词缀切分时,其词干出现歧义
- atalmighan(没能射中)=at(词干,表示射击)+al(能动式)+ma(否定)+ghan(完成体形动词词缀)
- atalmighan(称呼)=ata(词干,表示称为)+l(被动语态)+ma(否定)+ghan(完成体形动词词缀)

(2) 对同一个词进行词干、词缀切分时,其词缀出现歧义
- yazsila(您写)=yaz(写)+sila(第二人称尊称,表示您)
- yazsila(他一旦写)=yaz(写)+sa(条件式,第三人称单(复)数)+la(语气词)

(3) 词缀与词干的一部分相似,且切分、不切分均有实际意思
- qaymaq(奶酪)
- qaymaq(头晕)=qay(头晕,词干)+maq(动名词词缀)

(4) 词干的一部分被看成是词缀,错误地切除
- tamaq(饭,词干)
- ta+maq(错误的切分)

(5) 词缀被少切分现象
- didim(我说了)=de(说,词干)+dim(第一人称单数,过去式词缀)
- didim(我说了)=did+im(错误的分离)

(6) 音变字母可以还原成不同的字母,而且都有实际意义
- ëtip=at(仍,词干)+ip
- ëtip=et(做,词干)+ip

(7) 部分合成词的还原复杂,需要对多个弱化、脱落的字母进行还原
- ekel(拿来)=al(拿,词干)+ip(p 型副动词)+kël(来)

(8) 部分动词需还原成两个词
- achiq(拿出来)=ëlip(拿,词干)+chiq(助动词)

显然,维吾尔语词法较复杂,需考虑多种现象.总的来讲,维吾尔语词法分析需要解决以下问题:

1) 对音变现象进行还原:对一个词中发生音变现象的字母进行还原,这是词干、词缀切分的前提;
2) 词干、词缀切分:对一个词的词干和各词缀进行切分,丰富的形态变化已说明此工作的必要性;
3) 标注:对词干、词缀进行分类并标注.

2 维吾尔语词法分析的有向图模型

首先,我们给出维吾尔语词法分析的形式化描述:

对于给定的维吾尔语词的序列:$U_{1:n}=U_1,\ldots,U_i,\ldots,U_n$

第 i 个词的词干、词缀切分形式为

$$U_i = S_i + A_{i_1} + \ldots + A_{i_m} \tag{1}$$

其中,S 表示词干,A 表示词缀.

第 i 个词的标注形式为

$$U_i / t = S_i / t_{s_i} + A_{i_1} / t_{ai_1} + \ldots + A_{i_m} / t_{ai_m} \tag{2}$$

其中,t 为标注.

整个词序列的标注形式为

$$S_1 / t_{s_1} + A_{1_1} / t_{a1_1} + \ldots + A_{1_{m1}} / t_{a1_{m1}} + \ldots + S_n / t_{s_n} + A_{n_1} / t_{an_1} + \ldots + A_{n_{mk}} / t_{an_{mk}} \tag{3}$$

维吾尔语中每个词的构形一般由一个词干和一个或多个词缀的连接而构成,其制约关系不仅体现在词干、词缀之间,也体现在各词缀之间,如图 1 所示.ket 是词干,整个词的含义主要体现在词干上;-iwat 和-qan 都是词缀,

如:-iwat 是进行体词缀,有 4 个变体(-wat,-iwat,-uwat,-üwat),表示动作的持续性;-qan 是形动词词缀,有 4 个变体(-ghan,-qan,-gen,ken),起修饰下一个词的作用.词干后接某个词缀的哪个变体,是根据语音和谐规律来决定的.

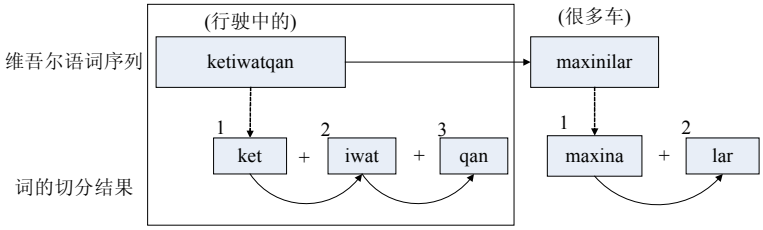


Fig.1 Relationship between the stem and affixes in Uyghur

图 1 维吾尔语词中词干与词缀的关系

从图 1 中可以看出,维吾尔语词具有明显的层次结构,而这种层次结构可以用树状结构来描述.这也是本文建模的依据.

2.1 词干、词缀切分的有向树模型

我们把维吾尔语中一个词的分析结果定义为链状层次结构,如图 2 所示.

图 2 中:S 表示词干(stem);$a_i(0\leqslant i\leqslant n)$表示词缀,其边表示词干到词缀、词缀到词缀之间的生成关系.对于整个词串,分析结果则可以描述为树状结构,如图 3 所示.

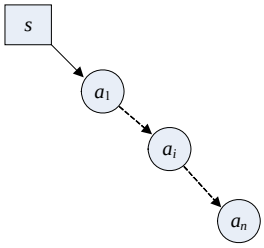


Fig.2 Directed Tree of a word

图 2 单词切分有向树

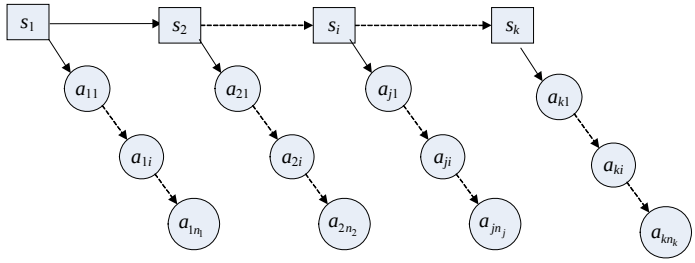


Fig.3 Directed Tree of word string

图 3 维吾尔语词串的有向树

与单个词的结构相比,整串分析结构中增加了相邻词干之间的转移关系,从而在所有词干和词缀之间形成一个拓扑有序的树结构.

模型中,我们给各节点之间的边赋予权重,用来衡量词干、词缀以及词缀与词缀之间的约束力度.那么,词法分析问题就转变为在所有的候选树中选择其概率之积为最高的树.模型中,我们将词干到词干的约束称为转移概率,将词干到词缀以及词缀到词缀之间的约束称为生成概率.

转移和生成概率的定义如下:

$$P(S_i|S_{i-1},\ldots,S_{i-n}),$$

其中,S 为词干,公式表示词干到词干的转移概率,类似于 n-gram 语言模型,i 表示词序,n 表示语言模型元数.

$$P(A_i|(S/A)_{i-1},\ldots,(S/A)_{i-n}),$$

其中,S 为词干,A 为词缀,S/A 表示词干或词缀,公式表示词干到词缀、词缀到词缀之间的生成概率.

给定一个候选树 T,我们用以上概率的乘积表示该候选树的整体概率:

$$P(T)=\prod_{S\in T}P(S\,|\,...)\times P(A\,|\,...) \tag{4}$$

为简洁起见,公式中隐藏了两个条件概率的历史条件.容易看出,这可以理解为传统的 n-gram 语法模型向树

结构的拓展.

2.2 切分与标注联合的有向图模型

上面的模型仅考虑词的词干与各词缀切分而没有涉及到标注.需要标注信息时,就必须同时对这些标注成分进行概率建模.

对于切分与标注相结合的模型,关键在于如何将标注信息有效地参与到词串中各词形态结构的生成过程中.本文工作中,对应于单纯切分的模型结构,我们为标注信息设计了一个同步树状结构,以描述词干和词缀标注之间的生成和转移关系.所谓同步是指标注树的结构和单纯切分模型的树结构完全一致,只不过树中对应节点,对前者而言是相应的标注,对后者而言是词干或词缀,如图 4 所示.

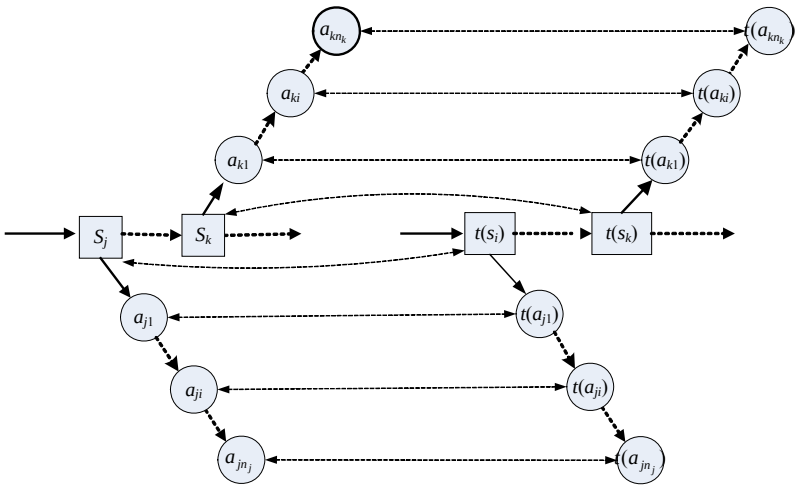


Fig.4 Directed graph model combined with segmentation and tagging

图 4 切分与标注相结合的有向图模型

模型中,我们又设计两项概率描述两个平行树结构中节点之间的映射关系:

$$P(X|t(X)),$$

X 代表词干或词缀,$t(X)$代表其标注.此项概率表示词缀 $t(X)$赋予词干或词缀的概率:

$$P(t(X)|X),$$

此项概率代表词干或词缀 X 被赋予标注 $t(X)$的概率.此项概率参与建模,使得模型倾向于选择常见的标注.

这两项条件概率在平行树结构的节点之间表示为具有双方向的边,从而建立起平行树结构之间的映射关系,目的是构建描述能力更强的有向图模型.

求解切分和标注的过程,即为在候选有向图中寻找概率积为最大的有向图.有向图 G 的概率定义为

$$P(G)=P(T)\times P(t(T))\times P(T,t(T)) \tag{5}$$

其中,$P(t(T))$表示标注树 $t(T)$的概率,它和 $P(T)$的定义一样,只需把词干和词缀换成相应的标注即可;$P(T,t(T))$表示平行树结构 T 和 $t(T)$的映射概率,它定义为平行树中所有节点对的条件概率的乘积,即

$$P(T,t(T))=\prod_{X\in T,t(X)\in t(T)} P(X\mid t(X))\times P(t(X)\mid X) \tag{6}$$

理论上,$P(G)$的 3 项乘子概率对于候选有向图的优选可能具有不同的决策力,所以应该为它们赋予合适的相对加权,有望提升模型性能.但在本文工作中,暂不考虑乘子加权问题,这相当于所有加权均为 1.

2.3 递归枚举词法分析候选

由于维吾尔语语音和谐规律的存在,枚举一个词的词法分析候选时需要对已发生音变现象的字母进行还

原,所以整个枚举候选过程可分为 3 个步骤:

(1) 枚举还原候选:根据词所包含字母是否有音变现象,枚举出可能对应的还原词;

(2) 对每个可能的还原词枚举出可能的词干、词缀切分候选;

(3) 从步骤(2)出来的每一项再递归枚举出其标注后的候选.

我们以 mektipining(译:他学校的)一词为例,描述词法分析过程,如图 5 所示.

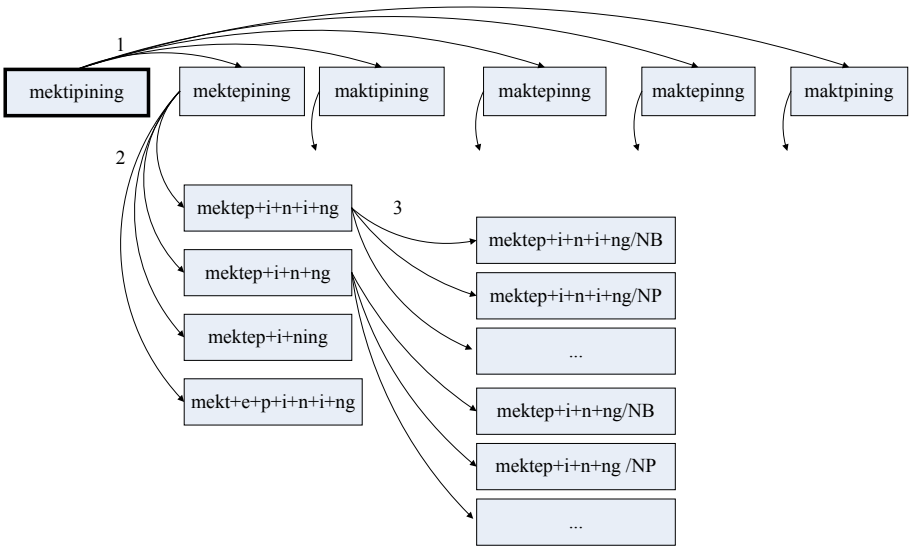


Fig.5 Recursive enumeration of instance candidates

图 5 递归枚举候选实例

首先,对 maktipining 一词枚举所有可能的还原形式(图 5 中由 1 来表示),对每一个候选再枚举可能的词干、词缀切分的形式(图上由 2 来表示),最后对每一个候选列出其可能的标注候选序(图 5 中由 3 来表示).枚举候选是递归过程,整个过程的算法见表 2.

Table 2 Algorithm of enumerate candidates recursively

表 2 递归枚举候选算法

Algorithm. Recursively enumerate candidates.

Input: A word ***w***;
output: Set of candidates {$\boldsymbol{S}+\boldsymbol{a}_1+\ldots+\boldsymbol{a}_n/\boldsymbol{t}$}.
1: //initialize vectors
2: ***V_restore***←0, ***V_sep***←0, ***V_tag***←0
3: //*V_restore* for fixed candidates
4: //*V_sep* for separated candidates
5: //*V_tag* for tagged candidates
6: add *RestoreFunc*(*w*)→***V_restore***
7: **for** *i*=0,…,*length*(***V_restore***) **do**
8: add ***Sep_Cand_Func***(0,***V_restore***[***i***],0)→***V_sep***
9: **endfor**
10: **for** *j*=0,…,*length*(***V_sep***) **do**
11: add ***tagFunc***(***V_sep***[***j***])→***V_tag***
12: **endfor**
13:
14: **function Sep_*Cand_Func***(*p*,*w*,*cur_pos*)
15: //*p* point to stem or suffix table, *w* is current word
16: //*cur_pos* point the current position of letter
17: **if** (*cur_pos*==*w.length*) **return** ***V_a***
18: //*V_a* is a vector, its construct has {*stemList*,*affixList*}
19: *symtable*=(*p*==0)?*stemtable*: *suffixtable*
20: **for** *n*=*cur_pos*,…,*w.length* **do**
21: *cand*=*w.substr*(*cur_pos*)
22: **if** (*cand* is in *symtable*)
23: **push** *cand* **into** ***V_a***;
24: ***Sep_Cand_Func***(*p*+1,*w*,*cur_pos*+1)
25: **else**
26: *cand*=*w.substr*(*cur_pos*+1)
27: **endfor**
28: **endfunction**

还原候选的枚举函数 *RestoreFunc* 是一个复杂的过程,也是维吾尔语词法分析的难点之一,本文将在下面的

篇幅给予详细介绍;词干、词缀切分候选函数 *Sep_Cand_Func* 的主要思路为:从给定的词中依次选取一定长度的字串,再与训练得到的词干、词缀表比较,决定是否为合法的词干或词缀,并将其词干、词缀压入相应的栈中,直到取完所有的字符串为止;标注函数 *tagFunc* 的功能与 *Sep_Cand_Func* 类似,这里不再解释.

根据枚举算法得到了所有的候选后,解码过程就是从候选列表中搜索概率之积为最大的路径.即,按公式(1)选择 $P(G)$值为最大者,此路径链接的树即为当前词串的词法分析结果.本文利用动态规划算法进行了解码,在此不再详述.

3 音变现象的自动还原模型

维吾尔语中孤立的词构造一个句子时往往会连接相应的词缀,如图 6 所示.

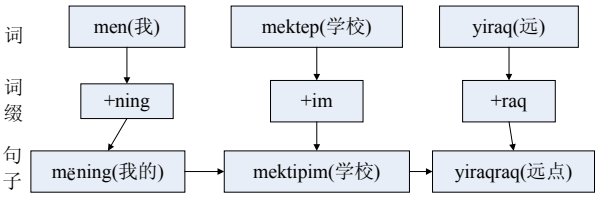


Fig.6 Procedure of Uyghur sentence generation

图 6 维吾尔语句子的生成过程

从图 6 可以看出,词干在接词缀时会发生音变的情况.所以,词干、词缀切分前需对发生音变现象的字母进行还原,它是正确切分词干、词缀的前提.首先定义几个术语:

术语 1(词的当前形式). 指词串中待分析的词,该词包含 n 个字母,其中,m 个($m \leqslant n$)字母发生了音变现象,用 w_{cur} 来表示.

术语 2(词的原始形式). 指发生音变现象之前的词,用 w_{org} 来表示.

如,对于 deptirim(我的本子)一词而言,w_{cur}=deptirim,w_{org}=depterim

针对音变现象,本文提出了还原模型:一个词所包含的字母构成线性序列,即 $w_{cur}=x_1x_2\ldots x_n$,假设其中 x_i 可还原为$\{\varnothing,y_{i1},y_{i2},\ldots,y_{im}|1\leqslant m\leqslant 32\}$($\varnothing$表示 x_i 是增音),并如果能找出所有字母可能原音集合,那么对给定词的还原过程,就相当于从每个字母众多的原音候选中找出最好的候选,此过程类似于线性序列标注问题.此模型的最大特点就是绕开音变现象繁琐的规则,通过建模加机器学习的方法来解决了还原问题.

还原模型的重点及难点是找出每一个字母可能的原音集合.我们从最小编辑距离算法的思路得到了启发,提出了词内字母对齐的算法.

3.1 词内字母对齐算法

词内字母对齐算法的目的是找出词内每个字母可能的原形集合.

设给定词的两种形式为 w_{cur} 和 w_{org} 且 $Len(w_{cur})=n$,$Len(w_{org})=m$,将 w_{cur} 和 w_{org} 所包含的字母尽可能一一对齐,找出一对同样的字母;对不能对齐的字母,则判断其为弱化、脱落或增音,并构成一个二元组:

$$\{\langle c_i,o_j\rangle|1\leqslant i\leqslant n,1\leqslant j\leqslant m\},$$

其中,c_i 为 w_{cur} 中第 i 个字母,o_j 为 c_i 可能的原形.

对 w_{cur},w_{org} 进行词内字母对齐时遇到的情况及对齐原则见表 3.

为了对齐的一致性,规定一个字母可以对齐到 $i(i=0,1,2)$个字母,i 的大小是经验值,是根据维吾尔语词的音变现象规律提出的.词内字母对齐算法关键是找到最佳的字母对,我们采用了动态规划算法实现,其主要思路见表 4.

算法中字母匹配函数是 *maxMatch*,其功能是依次比较 w_{cur},w_{org} 中的每个字母并匹配,查找对齐的字母对并压入栈.栈中每个项为一个字母当前形式与对应的原形.

Table 3 The principle of letter alignment in a word

表 3 词内字母对齐原则

	词内字母	对齐方案	对齐结果	备注
w_{org} w_{cur}	**a b c d** **a b e d**	a b c d a b e d	⟨**a**,**a**⟩, ⟨**b**,**b**⟩, ⟨**e**,**c**⟩, ⟨**d**,**d**⟩	弱化(c 弱化成 e)
w_{org} w_{cur}	**a b c d** **a b d**	a b c d a b d	⟨**a**,**a**⟩, ⟨**b**,{**b**,**c**}⟩, ⟨**d**,**d**⟩	脱落(字母 c 被脱落)
w_{org} w_{cur}	**a c d** **a b c d**	a ∅ c d a b c d	⟨**a**,**a**⟩, ⟨**b**,∅⟩, ⟨**c**,**c**⟩, ⟨**d**,**d**⟩	增音(b 是增音)
w_{org} w_{cur}	**a b c d** **a c d e**	a b c d ∅ a c d e	⟨**a**,{**a**,**b**}⟩, ⟨**c**,**c**⟩, ⟨**d**,**d**⟩, ⟨**e**,∅⟩	脱落、增音同时出现

Table 4 Algorithm of letter alignment in a word

表 4 词内字母对齐算法

Algorithm. Maximum matching.

Input: w_{cur}, w_{org};
output: ***V_pair*** //vector of pair letters.
push a (a∈w_{cur}) into ***V_cur***
push b (b∈w_{org}) into ***V_org***
add ***maxMatch***(***V_cur***,***V_org***)→***V_pair***

Function *maxMatch*(*V_cur*,*V_org*)
row=***V_cur***.*Len*; *col*=***V_org***.*Len*;
for *i*=0,…,*row*−1 **do**
  for *j*=0,…,*col*−1 **do**
pair: **if** *aligned*(***V_cur***[*i*],***V_org***[*j*]) **then**
    continue;
  else
    cur=*j*+*n* (*n*=0,1,2)
    if *aligned*(***V_cur***[*i*],***V_org***[*cur*]) **then**
      goto *pair*
    else
      push(***V_cur***[i],***V_org***[cur]) into ***V_pair***
    endif
  endif
endfor
endfor
return *V_pair*
endfunction

3.2 最大熵实例的抽取和训练

找出发生变音字母对后,即可利用机器学习的方法归纳出维吾尔语词中每个字母可能的原形候选.

本文利用最大熵模型对语料训练,训练语料库包括每个词的当前形式和人工还原的原始形式,对当前词内某个字母,选择其前后位置的字母作为特征模板,由表 5 给出.

Table 5 Feature templates of ME

表 5 最大熵特征模板

特征	意思	特征	意思
*C*0	当前位置的字母	*C*_2*C*_1	当前字母前面的两个字母
*C*1	当前字母的后一个字母	*C*_1*C*0	当前字母与前一个字母
*C*2	当前字母的后第二个字母	*C*0*C*1	当前字母与后一个字母
*C*_1	当前字母的前一个字母	*C*1*C*2	当前字母后面的两个字母
*C*_2	当前字母的前第二个字母	*C*_1*C*1	当前字母前后各一个字母

这里,我们使用张乐博士开发的最大熵工具包(http://homepages.inf.ed.ac.uk/lzhang10/maxent.html)进行最大熵训练,不使用高斯优先,只进行 100 轮迭代.

3.3 自动还原过程的解码

通过训练可得每个字母可能的原形候选.对给定的一个词进行还原的操作,实际上就是从每个字母众多的

原形后选中找出其概率积为最高的一个,如图 7 所示.

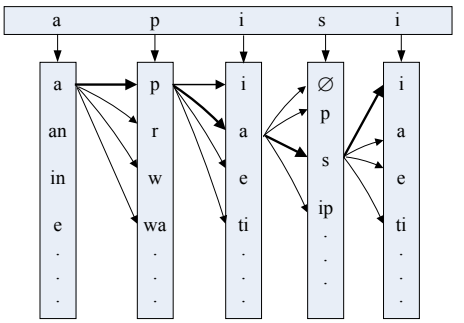


Fig.7 Schematic diagram of decoding process

图 7 解码过程示意图

通过解码,可以得到“apisi”(他的妈妈)一词的原始形式为“apasi”.

4 实 验

4.1 实验设置

我们在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》进行了实验.此语料库收集了小说《故乡》第 3 集,《棉花技术》、《小麦》等农业杂志的部分内容,《党的 17 大报告》、《知识—力量》、《新疆社科》等报告、杂志的部分内容,包括 67 114 个完整的句子,每个词条都有一级和二级标注,其中,一级标注数为 15 种,二级标注数为 71 种.

我们随机抽取各 5%的语句分别用做开发集和测试集,剩余 90%的语句用做训练集.模型各项概率从训练集中以极大似然估计(MEL)方法统计出.其中,词干到词干的转移概率、词干到词缀、词缀到词缀的生成概率以及相应标注之间的转移或生成概率,我们直接用成熟的语言模型工具 SRILM[32],以 WB 平滑方式训练 3 元模型.

维吾尔语的词法分析比较复杂,考察的方面较多,为了能够更好地体现系统性能,我们定义了多种指标,从不同的角度和层面测试了系统.首先给出考察指标的定义:

- 词干(S)级正确率 P_{stem}

以词干为单位,仅考察词干是否被正确提取、音变字母是否被正确还原,而不关心词缀的情况;

- 词级(W)准确率 P_{word}、召回率 R_{word} 和 F_{word} 值

以词为单位,仅当词内词干正确(包括音变字母的还原)、各词缀切分正确(包括音变字母的还原)时被认为正确分析,即:$w=S+a_1+\ldots+a_n$(S 为词干,a 为词缀)

4.2 实验及结果分析

实验 1:针对词干级、词级的实验.

实验目的:主要是测试系统对词干级分析、词级分析的性能.

实验安排:此实验安排在测试集上进行,为了更好地了解系统性能,同一个测试集上完成了两个实验,即:

(A) 不考虑标注,重点考察系统对词干提取以及整词分析的性能;

(B) 在实验(A)的要求上再把标注加到考察范围之内,即:不仅词干、词的分析正确,而且标注分析也需要正确,才被认为正确分析.

实验结果见表 6.

从表 6 的实验数据上可以看出,系统对维吾尔语的词干级(S)分析能力和词级(W)分析能力在不带标注时分别达到 94.7%和 91.6%,带一级标注时分别达到了 93.1%和 92.6%,带二级标注时达到了 89.1%和 89.0%.显然,系

统对带标注的分析能力比不带标注时的能力降低了几个百分点.分析原因是,因为维吾尔语中二级标注数量多,同时又存在同形异义词、兼类词,从而导致标注歧义.如:kokrek 一词表示“蓝一点”时做形容词,表示“胸部”时表示名词;kel(来)一词可以作为主动词(men keldim 我来了),也可作助动词(berip keldi 他去了),这需要足够的上下文信息才能分析得更正确,而系统中我们所考虑的上下文信息还有限,这也是我们后期改进模型中值得研究的问题之一.

Table 6 System performance on the test set

表 6 系统在测试集上的性能

性能	标注		
	实验(A)	实验(B)	
	–tag (%)	+tag (%)(一级标注)	+tag (%)(二级标注)
P_{stem}	94.7	93.1	89.1
P_{word}	93.1	93.4	89.9
R_{word}	90.0	91.8	88.3
F_{word}	91.6	92.6	89.0

实验 2:针对不同词性以及标注正确性的实验.

实验目的:主要测试系统针对某个指定词类的分析能力,除了测试词干级(*S*)、词级(*W*)分析能力之外,再测试制定词类标注(分别是一级 Tag_1 和二级 Tag_2)上的性能.

维吾尔语有 12 种词性,在句子中常见并起重要作用的主要是:名词(N)、动词(V)、形容词(A)、代词(P)和副词(D).我们在测试集上完成了此实验,实验数据见表 7.

Table 7 System performance on 5 different POSs

表 7 系统在 5 种词性上的性能

	V (%)				N (%)				A (%)				P (%)				D (%)			
	S	*W*	Tag_1	Tag_2	*S*	*W*	Tag_1	Tag_2	*S*	*W*	Tag_1	Tag_2	*S*	*W*	Tag_1	Tag_2	*S*	*W*	Tag_1	Tag_2
P	89.7	87.8	94.7	77.8	93.0	93.4	96.8	93.9	94.8	93.6	93.3	94.3	97.4	97.2	98.0	94.1	91.9	91.9	96.3	91.2
R	87.5	77.2	97.2	75.9	94.1	93.6	95.7	95.0	92.9	92.4	95.2	92.5	97.9	97.3	97.5	94.5	96.2	95.6	92.1	95.4
F	88.6	82.2	95.9	76.8	93.6	93.5	96.2	94.4	93.9	93.0	94.2	93.4	97.7	97.2	97.8	94.3	94.0	93.7	94.2	93.2

从表中数据可看出:所有的词性二级标注均低于一级标注,这个原因在实验 1 中已讨论过,在此不再讨论;所有的词性中,针对动词的分析不管是词干级还是词级分析,均低于其他词性.主要原因是:维吾尔语中动词是词缀数目最多、缀接层次最深、形态变化最丰富而复杂的词性,不仅词干中会出现与词缀相同的音节,而且各词缀之间有相似音节,如:*tamaq*(饭)与 *atmaq*(扔)=*at*(扔)+*maq*(词缀);*-ghan*(动名词词缀)与*-yidighan*(形动词词缀);这种相似性对于动词的分析带来歧义.通过什么方法能更有效地消歧,是维吾尔语词法分析中需要解决的另一个问题.

实验 3:针对自动还原模型的实验.

(1) 针对自动还原模型本身性能的实验

实验目的:测试自动还原模型对还原测试语料的性能.

实验安排:我们在开发集上完成了实验.

自动还原模型对开发集中整个词的还原正确率达到了 90%,其包括词干以及各词缀中发生音变现象字母的还原,说明本文提出的还原方法的性能在没有任何语言知识的情况下已经达到了满足实际需求的水平.此外,我们又分别测试了还原模块对不同词性的还原能力,实验数据见表 8.

Table 8 Performance of restore model

表 8 还原模块性能

	V (%)	N (%)	A (%)	P (%)	D (%)
P	71.0	91.2	98.1	97.6	92.1
R	69.2	91.0	96.3	99.1	97.3
F	70.1	91.1	97.1	98.3	94.6

从表 8 显示的数据可以看出,对于不同词性的还原性能,动词的还原率最低(70.1%),名词为倒数第二(91.1%).可以推测,动词还原率的低下是降低整个模块性能的主要原因.就像前面分析,维吾尔语动词的复杂变化特性是问题的关键.我们得到启发:能否通过引入一些语言规则的方法提高动词的还原率,从而提高整个模块的性能.

(2) 测试自动还原模块对整个词法分析系统性能的影响

实验目的:测试字母还原模块在整个词法分析系统中起到的作用.

实验安排:为了测试还原模块在整个系统中起到的作用,我们在开发集上分别进行两个实验:第 1 次将还原模块屏蔽掉(记为 Ex1);第 2 次将还原模块打开(记为 Ex2).两个实验以词级(*W*)分析作为测试指标,通过两个实验数据的比较得到还原模块在系统性能中到起到的作用以及幅度.图 8 为两次实验数据对比.

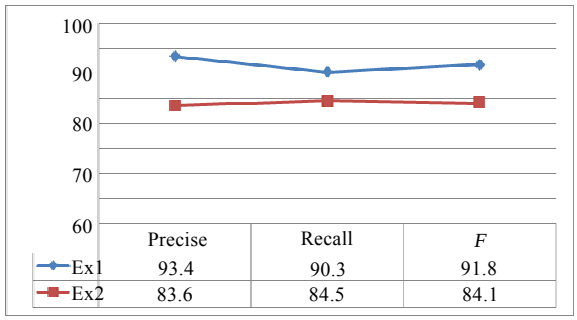


Fig.8 Effect of restore model on the performance of morphological analysis

图 8 自动还原模块对词法分析器性能的影响

测试结果表明,系统中加还原模块后,其 *F* 值达到 91.8%,比不加还原模块的 84.1%提高了 7.7%;准确率和召回率也都有相应的提高,说明自动还原模块在词法分析系统性能的提高方面确实起到了积极地的作用.

实验 4:针对语料库规模对系统性能影响的实验

实验目的:测试语料扩建会否提高还原模块的性能.

实验安排:我们固定开发集和测试集不变,而从训练集中每次提取不同规模的子集以训练系统,并考察该系统在测试集上的表现.整个训练集含 60 402 条句子,我们分别取训练集的 50%,20%,10%,6%,4%及 2%等不同规模的子集,并按照由小到大的次序对测试集进行词法分析.实验的评价标准是词级(*W*)*F* 值并带二级标注.图 9 为系统性能随训练规模增加的变化曲线.

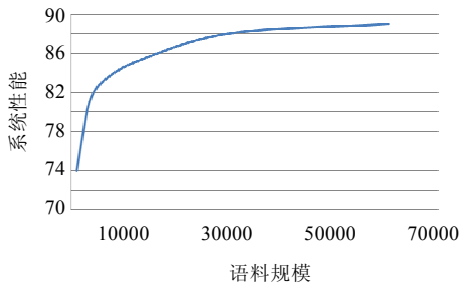


Fig.9 Curve between Training corpus size and system performance

图 9 训练语料规模与系统性能曲线

分析曲线可以发现,系统性能的提高的确与训练语料的规模有一定的关系,特别是语料规模较少的时候,这种依赖性尤为突出.比如,训练集由总规模的 2%扩大到 6%时,词法分析的正确率从 74%提高到了 81%.同时也可发现,随着语料规模的继续扩大,系统性能的提升幅度趋于缓慢.比如,训练语料规模由总规模的 50%提高到 90%

时,词法分析的正确率有88%提高到89%.这预示着训练语料的规模确实提高系统性能,但达到一定规模时,再继续扩大就失去意义.要提高系统性能:

(1) 可以扩大语料,特别是尽量收集不同领域、不同层面的语料,增强系统的适应性;

(2) 摸索将语言知识应用到系统的方法来增强模型.这也是我们今后系统性能的途径之一.

5 结 论

本文根据维吾尔语词的结构特征,提出了维吾尔语词的有向图模型,打破了将维吾尔语词串看成线性序列的常规.整体上,有向图由同步的词干、词缀树和标注树以及树间的映射关系组成,分别描述词干、词缀的生成、转移关系,相应标注的生成转移关系以及词干词缀与标注间的生成关系.通过转移或生成概率值作为这种关系的强度.同时,本文将有限字母的音变现象泛化到每一个字母上,将还原转化为类似于词性标注问题的方法,完全用统计的方法解决了字母的还原.不管是有向图模型还是还原模型都不依赖于任何规则,因此可以较方便地移到其他具有形态变化的黏着性语言上.

以后的研究目标为探索用判别式模式实现此模型,提高系统性能.我们目前只是根据从训练集中自动抽取出的词干表和词缀表为每个待分析词递归地穷举可能的候选结构,这导致过多的非法候选,以致引入无谓的歧义.如何利用语言学规则约束候选生成甚至解码过程,是我们未来要进行的重要研究内容.

References:

[1] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2): 257−286. [doi: 10.1109/5.18626]

[2] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging. In: Proc. of the Empirical Methods in Natural Language Processing Conf. 1996. 133−142.

[3] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th ICML. Massachusetts, 2001. 282−289.

[4] Adongbieke G, Abulimit M. Research on Uyghur word segmentation. Journal of Chinese Information Processing, 2004,18(6): 61−65 (in Chinese with English abstract). [doi: cnki:ISSN:1003-0077.0.2004-06-008]

[5] Wumaier A, Kadeer Z, Tursun P, Tian SW. Maximum entropy combined FSM stemming method for Uyghur. In: Proc. of the 2009 Oriental COCOSDA Int'l Conf. on Speech Database and Assessments. Urumqi, 2009. 51−55. [doi: 10.1109/ICSDA.2009.5278378]

[6] Kadeer Z, Wumaier A, Yibulayin T, Hamudula A. Uyghur noun inflectional suffix DFA generation. Journal of Chinese Infromation Processing, 2009,6(23):116−121 (in Chinese with English abstract). [doi: 10.1109/ICCSIT.2009.5234451]

[7] Kadeer Z, Yibulayin T. Uyghur adjective inflectional suffix FSM. Computer Knowledge and Technology, 2009,5(4):939−941 (in Chinese with English abstract). [doi: cnki:SUN:DNZS.0.2009-04-067]

[8] Aili M, Abulimiti M, Aimudula A. A morphological analysis based algorithm for Uyghur vowel weakening identification. Journal of Chinese Information Processing, 2008,22(4):43−47 (in Chinese with English abstract). [doi: CNKI:SUN:MESS.0.2008-04-006]

[9] Wumaier A, Yibulayin T, Kadeer Z. Noisy channel based Uyghur neutralized vowel identification model. Computer Engineering and Application, 2010,46(15):118−120 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2010.15.035]

[10] Ablimit M, Eli M, Kawahara T. Partly supervised Uyghur morpheme segmentation. In: Proc. of the Oriental-COCOSDA Workshop. 2008. 71−76.

[11] Aisha B. A letter tagging approach to Uyghur tokenization. In: Proc. of the Int'l Conf. on Asian Language Processing 2010. Harbin, 2010. 11−14. [doi: 10.1109/IALP.2010.72]

[12] Aisha B, Sun MS. A statistical method for Uyghur tokenization. In: Proc. of the Int'l Conf on NLP-KE 2009. Dalian, 2009. 1−5. [doi: 10.1109/NLPKE.2009.5313764]

[13] Oflazer K. Two-Level description of Turkish morphology. Literary and Linguistic Computing, 1994,9(2):137−148. [doi: 10.1093/llc/9.2.137]

[14] Larkey LS, Ballesteros L, Connell ME. Improving stemming for arabic information retrieval: Light stemmming and cooccurrence analysis. In: Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere, 2002. 275−282. [doi: 10.1145/564376.564425]

[15] McClosky D, Charniak E, Johnson M. Reranking and self-training for parser adaptation. In: Proc. of the ACL 2006. Sydney, 2006. 337−344. [doi: 10.3115/1220175.1220218]

[16] Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. Computational Linguistics, 1996,22(1):39−71.

[17] Zhang Y, Clark S. Chinese segmentation with a word-based perceptron algorithm. In: Proc. of the ACL. Prague, 2007. 840−847.

[18] Chiang D. Hierarchical phrase-based translation. Computational Linguistics, 2007,33(2):201−228. [doi: 10.1162/coli.2007.33.2.201]

[19] Jiang WB, Huang L, Liu Q, Lü YJ. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In: Proc. of the 46th ACL. Columbus, 2008. 897−904.

[20] Mi HT, Xiong DY, Liu Q. Research on strategies for integrating Chinese lexical analysis and parsing. Journal of Chinese Information Processing, 2008,22(2):10−17 (in Chinese with English abstract).

[21] Song Y, Cai DF, Zhang GP, Zhao H. Approach to Chinese word segmentation based on character-word joint decoding. Journal of Software, 2009,20(9):2366−2375 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3606.htm [doi: 10.3724/SP.J.1001.2009.03606]

[22] Zhu CH, Zhao TJ, Zheng DQ. Joint Chinese word segmentation and POS tagging system with undirected graphical models. Journal of Electronics & Information Technology, 2010,32(3):700−704 (in Chinese with English abstract). [doi: 10.3724/SP.J.1146.2009.00214]

[23] Bisazza A, Federico M. Morphological pre-processing for Turkish to English statistical machine translation. In: Proc. of the IWSLT 2009. Tokyo, 2009. 129−135.

[24] Mermer C, Akın AA. Unsupervised search for the optimal segmentation for statistical machine translation. In: Proc. of the ACL Student Research Workshop. Uppsala, 2010. 31−36.

[25] Eryiğit G, Adalı E. An affix stripping morphological analyzer for Turkish. In: Proc. of the IASTED Int'l Conf. on Artificial Intelligence and Applications. 2004. 299−304.

[26] Tai SY, Ong CS, Abdullah NA. On designing an automated Malaysian stemmer for the malay language (poster). In: Proc. of the 5th Int'l Workshop on Information Retrieval with Asian Languages. Hong Kong, 2000. 207−208. [doi: 10.1145/355214.355247]

[27] Tohuti L. The possibility of handling phonetic harmony by computer in Uyghur. Journal of the Central University of Nationality, 2004,31(5):108−113 (in Chinese with English abstract). [doi: cnki:ISSN:1005-8575.0.2004-05-022]

[28] Tomur H. Modern Uighur Grammar. Beijing: National Publishing House, 1987 (in Uighur).

[29] Wumaier A. Research on Uyghur morphological analyzing and syntax parsing key technology [Ph.D. Thesis]. Urumqi: Xinjiang University, 2010 (in Chinese with English abstract).

[30] Stolcke A. SRILM—An extensible language modeling toolkit. In: Proc. of the Int'l Conf. on Spoken Language Processing, Vol.2. Denver, 2002. 901−904.

附中文参考文献:

[4] 古丽拉•阿东别克,米吉提•阿布力米提.维吾尔语词切分方法初探.中文信息学报,2004,18(6):61−65.

[6] 早克热•卡德尔,艾山•吾买尔,吐尔根•依布拉音,艾斯卡尔•艾木都拉.维吾尔语名词构形词缀有限状态自动机的构造.中文信息学报,2009,6(23):116−121.

[7] 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自动机.电脑知识与技术,2009,5(4):939−941.

[8] 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.中文信息学报,2008,22(4):43−47.

[9] 艾山•吾买尔,吐尔根•伊布拉音,早克热•卡德尔.基于噪声信道的维吾尔语央音原音识别模型.计算机工程与应用,2010,46(15):118−120.

[20] 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.中文信息学报,2008,22(2):10−17.

[21] 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.软件学报,2009,20(9):2366−2375. http://www.jos.org.cn/1000-9825/3606.htm [doi: 10.3724/ SP.J.1001.2009.03606]

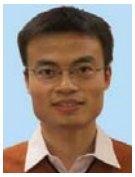
[22] 朱聪慧,赵铁军,郑德权.基于无向图序列标注模型的中文分词词性标注一体化系统.电子与信息学报,2010,32(3):700−704.

[27] 力提甫•托乎提.电脑处理维吾尔语语音和谐律的可能性.中央民族大学学报,2004,31(5):108−113.

[29] 艾山•吾买尔.维吾尔语词法分析句法分析中若干关键技术研究[博士学位论文].乌鲁木齐:新疆大学,2010.

麦热哈巴•艾力(1973−),女,新疆乌鲁木齐人,博士生,讲师,CCF 会员,主要研究领域为自然语言处理,机器翻译.

姜文斌(1984−),男,博士生,主要研究领域为词法分析,句法分析,机器翻译.



王志洋(1984−),男,博士生,主要研究领域为自然语言处理,机器翻译.

吐尔根•依布拉音(1958−),男,教授,博士生导师,CCF 会员,主要研究领域为自然语言处理,软件工程.

刘群(1966−),男,博士,研究员,博士生导师,CCF 会员,主要研究领域为机器翻译,自然语言处理.