文章编号：1003-0077（2017）00-0000-00

THUUyMorph：维吾尔语形态分析语料库*

* **收稿日期：** **定稿日期：**

基金项目： 国家自然科学基金重点项目(61331013)；国家“八六三”高技术项目(2015AA015407)

作者简介：哈里旦木（1978—），女，博士研究生，主要研究领域为自然语言处理；

哈里旦木·阿布都克里木[1]，阿布都克力木•阿布力孜[2]，孙茂松[1]，刘洋[1]

(1.清华大学 计算机科学与技术系，智能技术与系统国家重点实验室，清华信息科学与技术国家实验室（筹），北京 100084；

2.清华大学人文学院，计算语言学实验室，北京 100084）

摘要：该文介绍了维吾尔语形态分析语料库及其构建过程。我们从网上搜集了新闻、科技、小说、散文、日常用语和其它等不同领域的语料，采用制定切分规则（带语音变化和不带语音变化）、人工切分、错误分析和校对等过程建立了维吾尔语形态分析语料库。该语料库为 50 万词次规模，分为词级和句子级两类标注。该文工作不仅对相关维吾尔语语料库的建设具有参考意义，而且为维吾尔语的自然语言处理的研究提供了有益的资源。

关键词：THUUyMorph；维吾尔语； 形态分析

中图分类号：TP391 **文献标识码：A**

THUUyMorph: A Uyghur Morphological Analysis Corpus

Halidanmu Abudukelimu[1], Abudukelimu Abulizi[2], SUN Maosong[1], LIU Yang[1],

(1.State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;

2.Laboratory of Computational Linguistics, School of Humanities, Tsinghua University, Beijing 100084, China)

Abstract：This paper introduces the Uyghur morphological analysis corpus and its construction process. We have collected news, science and technology, novels, prose, daily language and other language materials in different fields, using the process of making the segmentation rules (with voice change and no voice change), artificial segmentation, error analysis and proofreading, the establishment of Uyghur morphology analysis corpus. The corpus is a 50 000 word scale, which is divided into two categories: word level and sentence level. The work of this paper not only provides a reference for the construction of relevant Uyghur corpus, but also provides a useful resource for the study of Uyghur natural language processing.

Key words: THUUyMorph; Uyghur; Morphology Analysis

1 引言

“一带一路”国家重大战略涉及 60 余个国家、50 余种国家通用语言和 200 余种民族语言，覆盖 44 亿人口，语言屏障问题逐渐成为推动深度国际合作与交流的重要阻碍。“一带一路”所涉及的绝大多数语言都是形态丰富语言。与汉语和英语等孤立语和屈折语不同，以维吾尔语为代表的形态丰富语言通过词干和词缀多种组合在词汇层面表示丰富的句法和语义关系，因而给语言技术处理技术带来严重的数据稀疏问题[1]。目前，国内面向形态丰富语言的信息处理技术相对而言严重滞

后于汉语和英语，远远无法满足“一带一路”语言互通的战略需求。维吾尔语形态分析是自然语言处理任务的基础性工作，为维吾尔语搜索引擎、机器翻译、信息检索、命名实体识别、句法分析、互联网舆情分析提供技术支持。因此，深入开展以维吾尔语为代表的形态丰富语言的处理技术研究具有重要的意义。

近年来，深度学习在自然语言处理中获得了成功的应用。相比传统基于统计的方法，深度学习不仅能够通过采用连续表示的方式缓解数据稀疏问题，而且能够自动从数据中学习特征表示，缓解了人工特征设计难以保证覆盖面的问题。然而，当前的深度学习技术主要是有监督的学习。因此，深度学习的成功运用首先要满足具有一定规模的标注语料[16]。

维吾尔语在语料库建立方面已有大量的工作。新疆大学吐尔根•依布拉音等[3-5] 和新疆师范大学的玉素甫•艾白都拉等[6-7]都已构建了百万词次的维吾尔语词法分析语料库，并分别在这些语料库基础上进一步进行了标注，如：词法及句法，以及面向具体任务的标注等。除此之外，文[8] 构建了FrameNet。文[10]建立了语法信息词典，文[11]建立了小规模命名实体关系语料库。虽然当前已有了相当规模的维吾尔语语料库[3-11]，但是还没有可公开的可供使用的维吾尔语形态分析语料库。

我们建立的公开形态分析评测语料库－THUUyMorph，是分为句子级和词级两种，专门用于维吾尔语形态分析有监督、半监督、无监督任务。 建立过程中我们参考了文本文工作建立和公开的维吾尔语形态分析评测语料库，不仅对维吾尔语形态分析的进一步研究，而且对整个维吾尔语自然语言处理任务也具有重要的参考意义。

2 研究背景

2.1 维吾尔语形态分析的特点

维吾尔语属于阿尔泰语系。阿尔泰语系有日语、韩语、朝鲜语、芬兰语、土耳其语、蒙古语和哈萨克语等 30 多种语言。维吾尔语在形态结构上属于黏着语类型的语言，作为黏着语言，词的词汇变化和各种语法变化都是通过在实词词干上缀接各种附加成分的方式来体现[12]。因此维吾尔语形态的多变性是维吾尔语的最突出的特点之一。

维吾尔语与英语、汉语有很大的区别：汉语词之间没有空格，分词后语法也被切开，而维尔语的词与词之间有空格，语法成分包含在词里。维吾尔语和英语的共同点是有前缀、词干和后缀的概念，但不同点是英语中前缀或后缀的个数一般为 1，而维吾尔语词的形态是词干后加若干后缀构成，而缀的个数通常超过一个。因此，英语、汉语等成熟语言的方法很难直接用在维吾尔语上。

2.2 维吾尔语形态分析的难点

维吾尔语形态切分是维吾尔语自然语言处理的一大难点，导致维吾尔语分词精度不高的原因一般有：黏着语、语音变化现象、歧义和形态切分问题等。

2.2.1 黏着语

维吾尔语作为一种黏着语在语素的组合上具有高度的灵活性，所谓的黏着性指的是维吾尔语的绝大部分附加成分都依附在词根之后，在同一个词根上依次地连缀几个附加成分，形成一种线条性特点[12]。虽然词干和词缀的数量有限，但是理论上可以组合生成无限的词语,其中，绝大多数维吾尔语词语在语料库中只出现一次[14-15]。维吾尔语通过在词干上添加词缀来实现丰富的句法和语义功能。以表 1 为例，“باغچىلاشتۇرۇماقچىمىز”（即“我们准备建园林”），作为一个词汇却表示出通常意义下句子层面的信息，这种情况在维吾尔语自然语言处理中造成了严重的数据稀疏问题。

表 1 维吾尔语的黏着性举例

维吾尔语词	汉语译文
باغ	园
باغچا	花园
باغچىلاش	花园化
باغچىلاشتۇر	使花园化
باغچىلاشتۇرماق	建园林
باغچىلاشتۇرماقچى	准备建园林
باغچىلاشتۇرۇماقچىمىز	我们准备建园林

2.2.2 语音变化现象

维吾尔语词缀种类多、数目多，而且可以多层缀接，在缀接过程中由于语音和谐规律某些语言会发生弱化、增音、脱落等音变现象，构成同一个词干的多种不同形态。实例见表 2.

从表 2 可以得出，维吾尔语的语音和谐规律不是简单的结合，因此，建立带有语音和谐规律的形态分析语料对维吾尔语的自然语言处理研究具有重要的意义。

表 2 维吾尔语语音变化现象

音变现象	例子
弱化	**مەكتەپ**（学校）+**م**（第一人称单数）= **مەكتىپىم**（我的学校）
增音	**ئارزۇ**（愿望）+**يۇم**（第一人称单数）=**ئارزۇيۇم**（我的愿望）
脱落	**بۇرۇن**（鼻子，词干）+**ى**（第三人称单数，词缀）=**بۇرۇنى**（他的鼻子）
多种现象同时出现	**قال**（留）+ **ىپ** + **تۇ** +**كەن**（系助动词） = **قاپتىكەن**（听说他留下了）

2.2.3 歧义

维吾尔语词的歧义想象也较严重，例如تۇردى这个词在下面的两个例子中分别表示不同的意思：

1）原文：.تۇردى كەلدى

意义 来 吐尔迪。

释义：吐尔迪来了。

2）原文 تۇردى ئۇ

意义:站 他

释义：他站起来了。

这个例子تۇردى 也表示“吐尔迪”（人名），也表示动词“站”的意思。除此之外维吾尔语中还有其它词类也存在歧义现象。n 例如：بارماق 表示手指，但是 بارماقتوشقان这个词也表示“去”，توشقان 表示“兔子”，但是توشقانئالما也表示“满，丰满”，ئالما 表示“苹果”ئالما 又表示动词“拿的”否定形式“别拿，不要拿”。يېڭى 表示“新”，但是它还表示“袖子，يەڭ”附加人称“ى”时发生语音变化“يېڭى，袖子”等等。维吾尔语的兼类词现象比较严重，这种现象对维吾尔语形态分析任务带来一定程度困难。

2.2.4 形态切分问题

目前情况来看维吾尔语的形态划分问题上，还存着意见分歧。传统形态学把形态变化的附加成分，分为构词附加成分（构词词缀）和构形附加成分（构形词缀）。构词词缀，功能是构成新词。构形词缀是不会改变词义，而只改变词的语法意义并表示词的各种语法关系。有的维吾尔语语法书把构词词缀称为“词缀”，则把构形词缀称为“词尾”[13]。上述分类方法有很多不足之处，自相矛盾的地方。维吾尔语里面的一部分附加成分，从形式上看，他们好像是构词附加词缀，但是其功能上他们却具有构形附加成分的功能。如：بىلىم+دىم+م (我的知识)，这里有两个-م,形式相同，但功能不一样，第一个-م是构词词缀，使 بىل -（知道）动词改بىلىم(知识)名词。第二个-م是构形词缀，它只是第一人称单数词缀，也不变بىلىم（知识）的意义，也不变词性，指名词属于第一人称。因此对这些词汇进行自动形态分析很难达到预期目标。

3 维吾尔语形态分析标注库建设

3. 1 标注规范

3.1.1 基本规则

维吾尔语形态分析标注规范由总体原则、切分方法、根据不同的词性详细的介绍了词干和词缀的切分规则，并给出了相应的实例，为了能够提高阅读的通用性，整个规范文档由中文书写，维吾尔语文部分采用了对应的拉丁字母的书写形式，同时在文档里给出了维吾尔文字母对应的拉丁字母 。

词干是一个词除去构形附加成分的部分。词干可能是由词根构成的，也可以是词根加上构词附加成分的。例如：يازغۇچىلار （作者）其中-لار 是词尾，ياز - 是词根，- غۇچى 是构词附加成分，这个词除去构形附加成分，剩下的يازغۇچى 就是词干。

1）维吾尔有两种词缀：构词词缀和构形词缀，本文只考虑构形词缀的形态切分。例如：

ساياھەت 旅游

ساياھەتچى 旅游者

ساياھەتچى#نىڭ 旅游者#的

ساياھەتچىلىك#نىڭ 旅游业#的

本论文中旅游者、旅游业属于构词词缀，而旅游者的、旅游业的属于构形词缀，本论文的形态切分任务是将“旅游者的”和“旅游业的”分别切分成“旅游者#的”和“旅游业#的”，而构词成分“旅游者”和“旅游业”不切分。

2）当词干单独出现时，不加任何标志，默认为词干。如：旅游。

3）当词干与构形词缀一起出现时，词干后面“#”与词缀分开。如：旅游者#的。

4）当词干或词缀发生语音变化时，后面加小括号，括号内部写原形。如：ئائىلى.(ئائىلە) لەر(لار) ى مۇ

3.1.2 切分细则

我们主要是从名词，形容词，数词，量词，副词，代词，动词为依据来切分。目前进行的是粗切分，即构形切分。

1) **名词：**名词原型（名词的主格形式）为词干，派生名词(名词的零派生形式)、专用名词可以单独做词干，比如，人名，名词后面加各种名词人称、格、数语法范畴时，名词语法范畴和名词词干分开。
2) **形容词：**形容词的原形和最高级被认为词干（维吾尔语形容词的最高级不带任何构形词缀），但是形容词的减少级和亲热级要切分，但是形容词的部分减少级被认为词干,قىزغۇچ（浅红色）.
3) **数词：**数词跟其它成分分开，基数是数词词干，数词的其它形式要切分。如,تۆتىنچى（第四），ئالتىمىز（我们六个）等。
4) **量词：**量词跟其它成分翻开，量词没有加构形附加成分的部分就是量词词干，当量词后面加词缀时词缀和词干要分开。كىلومېتىرغا (每公里)等。
5) **副词：**维吾尔语中大部分副词时独立出现作为词干来处理，很少一部分副词是带后缀， 后缀和词干部分要分开。如，تېزراق（快点），هازىرغىچە（直到现在）
6) **代词：**代词单数是代词词干，代词复数要切分，除此之外维吾尔语代词经常与名词词缀组合，因此这种形式的代词要与词缀分开。如，سىزنى（把你），بىزنىڭ （我们的）等。
7) **动词：**动词带静词化附加成分及时态附加成分的部分是动词词干。因此动词带的语态附加成分、体语附加成分否定附加成分、静词化附加成分、时态附加成分、人称附加成分、语气语附加成分都与词干切分。如：ماڭدىم（我走了），يازدىم（我写了），ئالىمىز （我们要买），تىرىشىپ （努力） 等等。
8) **模拟词：**单独出现的模拟词，即所有的模拟词是词干。
9) **连词：**连词单独出现的形式是词干，附带实词作构形附加成分的形式要切分。如，ھەم（和） , ياكى （或） ,كەل+دى+يۇ （大概意思“就”的连词）等。
10) **后置词：**后置词本身就是词干形式出现。
11) **语气词：**单独使用的语气词本身被视为词干，附带实词作构形附加成分的语气词要切分。如：بەلكىم （可能）سەن+چۇ（你呢） كەل+دى+غۇ（他也来了呢） 等。
12) **感叹词：**维吾尔语中的所有感叹词以词干形式出现。

除此之外，维吾尔语中的缩略词基本上存在三种情况：

1) 只取每个词的首字母，并用空格隔开，因此目前不存在切分问题。如：ج ك پ, ب د ت。
2) 取第一个词的第一个音节和最后一个词的音阶，合并成为一个词干。如：پارتكوم-پارتىيە كومپىتتى 等。
3) 用拉丁字母缩写，没有包含在维吾尔文语料中。如：GDP, WTO,KTW 等。

3.2 形态分析语料库建立流程

我们首先从天山网（http://uy.ts.cn/）、新华网（http://www.xinhuanet.com/）、Misranim（Misranim.com)、Alkuyi(Alkuyi.com)、Alimahat（Alimahat.com）、Tarimweb（tarimweb.com）和 Anatuprak（anatuprak.com）等网站上搜集了涉及新闻、科技、小说、散文、BBS 和其它等不同领域的语料，总的词语数量为 575 103 个，词语类型数量为 118 852 个，不同领域词语数量的具体分布如图 3 所示：

表 3 不同领域词语数量的具体分布

领域	词语数量	词语类型数量
新闻	200 775	21 147
科技	60 120	15 096
小说	122 367	27 503
散文	56 698	18 967
BBS	80 074	20 146
其它	55 069	15 993

其次，我们使用 tokenizer.perl（https://github.com/moses-smt/mosesdecoder）工具对语料进行了标点符号切分的预处理。同时，为了减轻标注的工作量，我们提取了语料中的词语类型作为人工标注的数据。我们从中央民族大学维吾尔语语言学专业的学生中自愿选择了 7 位学生对语料进行人工的形态切分，要求对每一个词进行无语音变化，有无语音变化两种切分。在人工标注过程中不断对语料和人工切分错误及不一致性进行反复更正。人工标注完成后，从 7 位学生中选出标注最好的学生进行了 1 次校对，之后又邀请了新疆大学的阿布都热依木老师和这位学生交替的进行了 3 次校对，总共进行了 4 次校对，最终建立的词级和句子级的形态切分语料库。

3.4 维吾尔语语音变化现象分布

我们对人工形态切分后的新闻领域的语料的词表进行了语音变化想象的统计。新闻词表大小为 19621,其中发生语音和谐变化的词有 4688，占总词表的 23.9%。为了进一步了解发生语音和谐变化的 4688 个词中词干和词缀在不同语音和谐变化现象下的分布，我们做了统计，见表 4.

表 4 语音和谐变化现象分布

语素	弱化(%)	增音(%)	脱落(%)
词干	95.7	2.2	2.1
词缀	97.2	1.2	1.6

从表 8 可知，语音变化现象主要体现在弱化现象上，而且词干和词缀的分布相似。一般情况下，语音和谐变化发生在词干或语素内部，而语素之间不会发生语音和谐变化。因此我们得出，维吾尔语中语音和谐变化很严重，其中，弱化现象应该成为重要研究点。

3.5 维吾尔语词级形态分析语料库

我们将维吾尔语词级形态切分语料库有 19621 个维吾尔语词。我们分为训练集，开发集和测试集。训练集有 2 万条词表，开发集和测试集分别是 1000 条词表。测试任务分为两种，一种是只进行词干和词缀的切分，一种是词干，词缀切分的同时考虑语音变化。原始语料为从 2012 年-2013 年的天山网下载语料，进行词频统计，建立了词和对应的词频词表。语料库构建步骤分为：正字法检查，词干，词缀提取，语音变化。通过以上步骤建立了 2 万多词的形态分析词表，为了满足不同的用户，分类维吾尔文字，根据任务的不同，分为词干词缀切分词表和带语音变化的词表。然后通过四次校对提取单词以后进行标注。

3.5.1 不带语音变化的数据

标注数据分为未带语音变化和带语音变化。未带语音变化语料数据文件每一个行包括一个单词以及该该单词的切分，切分是对切分以后的词干不进行还原，即词干变形。单词和切分用 Tab 键分开，切分用空格键(Space)分开。如下例：

ئائىلىدىكى ئائىلى دىكى
ئائىلىدىكىلەر ئائىلى دىكى لەر
ئائىلىدىن ئائىلى دىن
ئائىلىدە ئائىلى دە
ئائىلىسى ئائىلى سى
ئائىلىسىدىكىلەر ئائىلى سى دىكى لەر
ئائىلىسىدىكىلەرنىڭ ئائىلى سى دىكى لەر نىڭ

3.5.2 带语音变化的数据

带语音变化语料库数据文件每一行包括一个单词以及该单词的切分，切分时对切分以后的单词进行还原。单词和切分用 Tab 键分开，切分用空格键(Space)分开。如下例：

ئائىلىدىكى ئائىلە دىكى
ئائىلىدىكىلەر ئائىلە دىكى لەر
ئائىلىدىن ئائىلە دىن
ئائىلىدە ئائىلە دە
ئائىلىسى ئائىلە سى
ئائىلىسىدىكىلەر ئائىلە سى دىكى لەر
ئائىلىسىدىكىلەرنىڭ ئائىلە سى دىكى لەر نىڭ

3.6 维吾尔语句子级形态分析语料库

我们进一步完善形态分析语料的建设，在词级语料库的基础上建立了句子级形态分析语料。句子级语料包含 15327 条句子。因为词级形态分析语料建设中已经建立了标注规范，词级规范直接应用到句子当中。句子级形态分析时句子中的每一个词进行人工形态切分并校对，词干和词缀之间＃号来分开，词缀和词缀之间与 / 号来分开。如下例所示。

مەزمۇن#/ى/نى تەكشۈر#/ۇش ئىش#/ى/نى

بەش خىزمەت كۇن#/ى/دە بېجىر#/پ

تۈگىت#(تۈگەت)/ىمىز.

句子级语料的建设比词级形态语料建设有几方面的优势，1）句子级形态分析时完全可以按上下文来判断句子中每一个词的词干部分，这样就避免兼类词难切分的情况。2）句子形态分析时可以避免一些正字法错误，方言词等词汇错误切分。

3.5.1 实验

我们对句子级形态分析语料库进行了统计，见表 5。

通过实验我们发现词、词干、缀的平均长度是 17、14 和 3.5。维吾尔语词的最大长度为 33。不管是词级还是句子级的形态分析语料库，都为维吾尔语的形态分析的研究工作提供的语料。目前，已有工作使用了维吾尔语形态分析词级未带语音变化的语料，即文[14] 使用该数据研究了维吾尔语形态切分在神经网络中的性能体现，获得了具有参考价值的实验结果。

表 5 维吾尔语形态切分新闻语料库统计

词的最长长度	33
词的最短长度	1
词的平均长度	17
词干的最长长度	27
词干的最短长度	1
词干的平均长度	14
缀的最长长度	9
缀的最短长度	1
缀的平均长度	5
词缀的最多个数	6
词缀的最少个数	1
词缀的平均个数	3.5
带词缀的词的比例	53%
词干发生变型的词的比例	12%
词缀发生变型的词的比例	3%
词干与词缀同时变型的比例	0.6%
只有 1 个缀变型的比例	2.6%
有 2 个或 2 个以上缀变形的比例	$8.263*10^{-5}$%

4 结论

本文描述了我们构建的维吾尔语形态分析语料库，并着重分析了维吾尔语形态分析的一些特点及其在形态切分方面所带来的困难。在该语料的构建方面，我们从网上搜集了涉及新闻、科技、小说、散文、BBS 和其它等不同领域的语料，采用制定切分规则，人工切分，错误分析和校对等过程建立了 50 万词次规模的维吾尔语形态分析语料库，计划公开该语料库。

参考文献

[1] Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (SPMRL) What, how and whither . In Proceedings of the NAACL Workshop on Statistical Parsing of Morphologically-Rich Languages . pages 1–12. http://www.aclweb.org/anthology/W10-1401 .

[2] Zohp B, Yuret D, May J, et al. Transfer Learning for Low-Resource Neural Machine Translation [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, 2016: 1568-1575.

[3] 吐尔根•依布拉音, 阿里甫•库尔班. 基于词典的现代维吾尔语词性自动标注系统的研究[C] // 中文输入技术发展历程及输入方案汇编（论文集）, 2006, 11.

[4] 艾山•吾买尔. 维吾尔语词法句法分析器中的词义排歧问题的研究[D][博士学位论文]. 新疆大学. 2009.

[5] 买合木提•买买提, 吐尔根•依布拉音. 基于 n-gram 的维吾尔语词性标注研究[C] // 第二届中国少数民族青年自然语言处理学术研讨会. 2008, 10:185-189.

[6] Y. Aibaidula and K. T. Lua, "The Development of Tagged Uyghur Corpus," in 17th Pacific Asia Conference on Language, Information and Computation, D. H. L. K. T. Ji, Ed., 2003, 228-234.

[7] Y. Abaidulla, 1. Osman, and M. Tursun, "Progress on Construction Technology of Uyghur Knowledge Base," in

2009 International Symposium on Intelligent Ub iquitous Computing and Education" 2009, 554-557.

[8] R. Mirejiguli, K. Alifu. Design of the Uyghur FrameNet Desktop. Singapore 3.1. (Feb 2015)

[9] K. Abiderexiti, M. Maimaiti, A. Wumaier, and T. Yibulayin, "Construction of Uyghur Initial Paraphrase Corpus," in Proceeings of the International conference "T'urkic Languages Processing" T'urkLang 2015, Russia,Tatarstan,Kazan, 2015, 87-90.

[10] J. Wushouer, W. Abulizi, K. Abiderexiti, T. Yibulayin, M. Aili, and S. Maimaitimin, "Building Contemporary Uyghur Grammatical Information Dictionary," in Worldwide Language Service Infrastructure, Y. Murakami and D. Lin, Eds. Cham: Springer International Publishing, 2016, 137-144.

[11] K. Abiderexiti, M. Maimaiti, and T. Yibulayin, et al., Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus. 2016 Internation Conference on Asian Language Processing (IALP). 103-107.

[12] 艾孜尔古丽，阿力木•木拉提,玉素甫•艾白都拉.基于形态分析的现代维吾尔语名词词干识别研究. 中文信息学报，2015 年，第 29 卷 6 期

[13] 霍盛，试论维吾尔语形态变化的功能及特点。新疆大学学报（哲学社会科学版），1991 年 19 卷第三期

[14] 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 [J]. 清华大学学报：（自然科学版）, 2017, **57**(1): 1-6.
ABUDUKELIMU Halidanmu, CHENG Yong, LIU Yang, et al. Uyghur morphological segmentation with bidirectional GRU neural networks [J]. *J Tsinghua Univ*: (*Sci and Tech*), 2017, **57**(1): 1-6. （in Chinese）

[15] Abudukelimu H, Liu Y, Chen X, et al. Learning Distributed Representations of Uyghur Words and Morphemes [C]// Proceedings of CCL/NLP-NABD. Guangzhou, China, 2015: 202-211.

[16] Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting Finite-State Transductions with Neural Context. In Proc. of NAACL .

作者联系方式：

姓名：哈里旦木·阿布都克里木

地址：北京市海淀区清华大学 FIT 楼 4-506

邮编：100084

电话：18811365418

电子邮箱：abdklmhldm@gmail.com