# 目的：

## 解决什么问题

本文是主要研究该项目的维吾尔人名识别模块，它是维吾尔文命名实体识别中难度较大而且特别重要部分。

（备注：本论文举例的维吾尔文单词、词组、句子都是用标准的维吾尔拉丁文表示的。）

众所周知，与其它语言的命名实体识别研究成果相比，维吾尔人名识别还是

处于一个初步的研究阶段。维吾尔文命名实体识别是维吾尔文信息处理任务的前

提和基础，其中人名占有相当大的比例，而且是最难识别的一部分。其识别精度

还远远达不到实际应用的需要。

1.绪论:

综述课题来源, 介绍背景和研究意义,

详述命名实体识别的国内外研究现状，

最后讲述维吾尔语命名实体识别的研究现状以及获得的一些科研成果。

2.知识库和前期准备工作：

维吾尔人名的结构特性（ok）和难点

综述识别维吾尔人名过程中所需的知识库的建立

知识库的结构

相关的一些前期研究工作

（概念--未登录词 即没有[被收](http://www.baike.com/sowiki/%E8%A2%AB%E6%94%B6?prd=content_doc_search)录在分词词表中但必须切分出来的词，包括各类专有名词（人名、地名、企业名等）、缩写词、新增词汇等等。）

3. 候选人名的提取

详细论述了（识别维吾尔人名当中所用到的）候选人名提取方法，

提取时所需的重要模板和统计算法

然后介绍了候选人名提取的具体步骤。

4. 特征提取和规则施加

描述维吾尔人名识别过程必要的特征集以及它的 具体提取方法 和 步骤，

然后讲述了提取的候选人名当中识别人名的规则及其特征和规则的施加步骤。

5. 消除歧义

主要介绍了歧义性维吾尔人名的结构特点，

并总结相应的词法和语法特征，

对具有歧义性的人名进行有效的消歧。

6. 系统模型及实验结果分析

综述了维吾尔人名识别系统的具体模型、识别流程、操作方法，

然后介绍了所进行的实验及其实验结果，

并详细介绍实验结果的分析。

7. 总结和展望

全面介绍了本文的重要内容，指示了将来进行的一些重要研究工作，以及对系统一些优化意见。

8.自己的一些想法

雏形：

1. 背景介绍

众所周知，与其它语言的命名实体识别研究成果相比，维吾尔语命名实体识别工作还在处于初步的研究阶段，维吾尔文命名实体识别是维吾尔文信息处理任务的前提和基础，其中人名占有相当大的比例，而且是最难识别的一部分。其识别精度还远远达不到实际应用的需要。

（缺少一段连接）

1. （维吾尔族人名介绍

维吾尔人名的结构跟外国人名比较相似，格式是“本名·父名”，即本名在先，父名在后。维吾尔人名和外国人名相同的是，本名和父名都是以空格分开的。如 Ismayil Emet

（司马义·艾买提），Ismayil“司马义”是本名，Emet“艾买提”是父名。

 维吾尔人名没有姓氏概念，起名方式以“本名+父名”格式延续，，而

且名字各种各样，种类特别多，但是在一定程度上可以归纳名字的种类：

（1）. 以日月星辰为名; 如 Shemis 谢木西（阿拉伯语），意

“太阳。Qemer 克买尔（阿拉伯语），意“月亮”

（2）以山川乡土为名；如Pamir qiz  帕米尔克孜，“帕米尔”即高原名

（3）以花卉果木为名

（4）以鸟兽之名为名；如：女孩有的命名为 Bulbul qiz 布里布力克孜(百灵鸟)

（5）以十二生肖为名；

（6）以里大事件命名；如解放初男孩命名为 Azat（阿扎特），意“解放”，女孩命名为 Azatgul（阿扎特古丽），  意“解放之花”。

）

3.维吾尔人名（格式是“本名·父名”）的数量很多，规律性差、随意性大、结构成分复杂、歧义性较大，识别维吾尔人名的过程会遇到以下难点：

1）人名长度不固定。一个单词或几个单词可以代表维吾尔人名。

2）人名词性复杂。

3） 机构名、地名中出现人名。

4）存在二义性。比如，Adil sinipimizning oqoghuchisi（阿迪力是我们班的学生）这个句子中 Adil（阿迪力）作为人名，Hemme ademge adil bolayli（我们要公平对待每个人）这个句子中 Adil（阿迪力）作动词。

5） 黏着性导致派生词。

6）人名写法不一致。同一个人名的不同写法出现在真实语料库中。比如：Hebibulla（艾比布拉）在语料库中写法很多，如，Hëbibilla、Hibibilla、Hëbilla、Hibilla、Hëbi、Hebuq、Hebi 等等。

3.相关技术、工具的选择（有问题）

1）基于统计和规则相结合的方法

在实际的命名实体识别工作当中，一般不会单独采用基于规则或者基于统计

的方法，比较常用的是统计和规则相结合的方法。先用基于统计的方法对输入的文本进行命名实体识别，然后再用基于规则的方法对识别结果进行进一步纠正和优化，以达到较高的准确率。

2）词干提取

识别人名过程中词干提取起着特别重要的作用，尤其是对黏着性语言词干提

取工作是必不可少的。

本文采用新疆大学多语种重点实验室研发的词干提取工具，解决了维吾尔人

名的词干提取问题。本词干提取工具是用规则和统计结合的方法提取词干，正确

率超过 95%，而且提取速度特别快。

3）维吾尔文分词

本文利用新疆大学多语种重点实验室研发的维吾尔文分词工具的动态链接库（DLL），此工具的正确率已到达 99%；

4）知识库

维吾尔人名库：本文采用新疆民族出版社 2006 年出版的“常用的维吾尔人名”书中的维吾尔人名，书中一共有 24703 个人名，其中13773 个是男人的名字，10930 个是女人的名字。

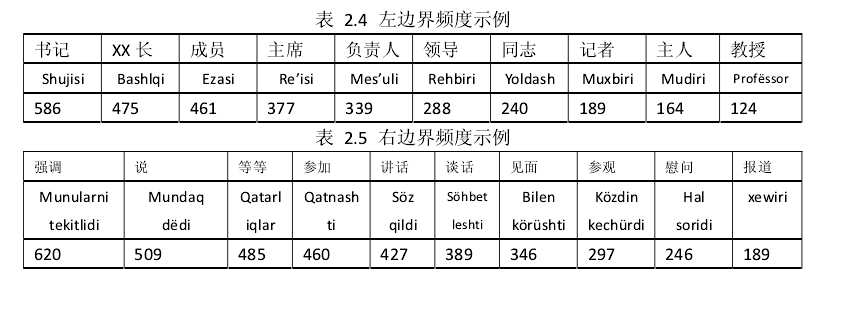
其他：著名人物的名字库、地名库、机构名库；

5）辅助语料

人名边界词知识库、人名前缀后缀词库、歧义性人名库 、称谓词库、地名特征词库、机构名特征词库。

其中关于人名前缀后缀词库，本文总结出人名前后经常出现的一些边界成分，我们把它称为左边界（LeftBound）和右边界（RightBound），其中 1281 个是左边界、1364

个是右边界。表 2.4 和表 2.5 给出了预料中出现频度最高的前 10 个左、右边界词



4.具体算法：（有问题）

基于边界模板为基础的局部统计的算法：

此算法所需要的模板和其它知识库可以从语料库中容易得出。

（该算法需要的语言资源可完全从标注语料库中自动抽取;）

1. 在利用边界模板对人名进行粗略定界以后,
2. 根据局部统计量和几条有效的启发式规则对识别结果进行校正;

(首先用边界模板对候选人名定界，然后利用局部统计量和一些识别规则对候选人名进行纠正。)

我们的算法针对候选人名提取做了如下几条假设：

（1） 由于维吾尔人名没有姓氏概念，因此完整的维吾尔人名是由自己的名字加

父名组成，一般不超过 3 个单词。

（2） 若某长度不小于 3 的字串在一篇文章中多次出现，且某次出现被识别为人

名，则可以将其所有出现的该字符串均识别为候选人名。

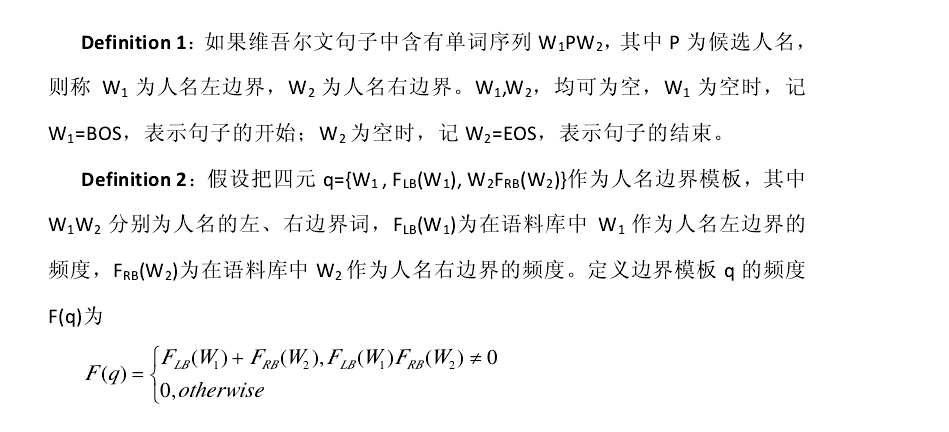
（3） 长度小于 3 个字符的单词不能作为候选人名。因为据我们的统计维吾尔人

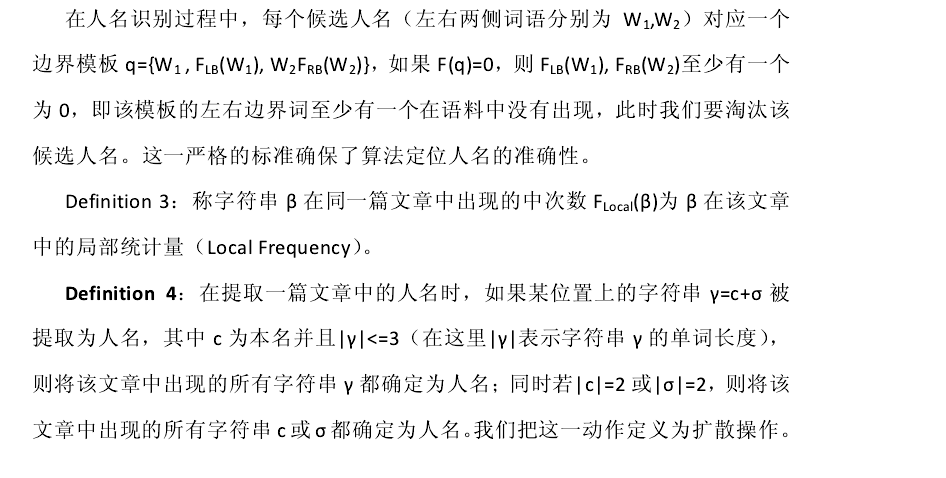
名当中长度最小的人名是 Eli（艾力），所以长度小于 3 个字符的单词不算

人名。

（4） 统计时以单词作为最小统计单位。由于维吾尔人名由单词组成，因此以单词统计是最有效而且可靠。

5.边界模板和统计量的定义





人名识别所需资源及其提取：

我们把“人民日报”、“天山网”和“新疆维吾尔作家协会”等网站的维吾尔

人名较多的 1602 片文章作为语料并进行了左右边界提取和统计。我们一共提取

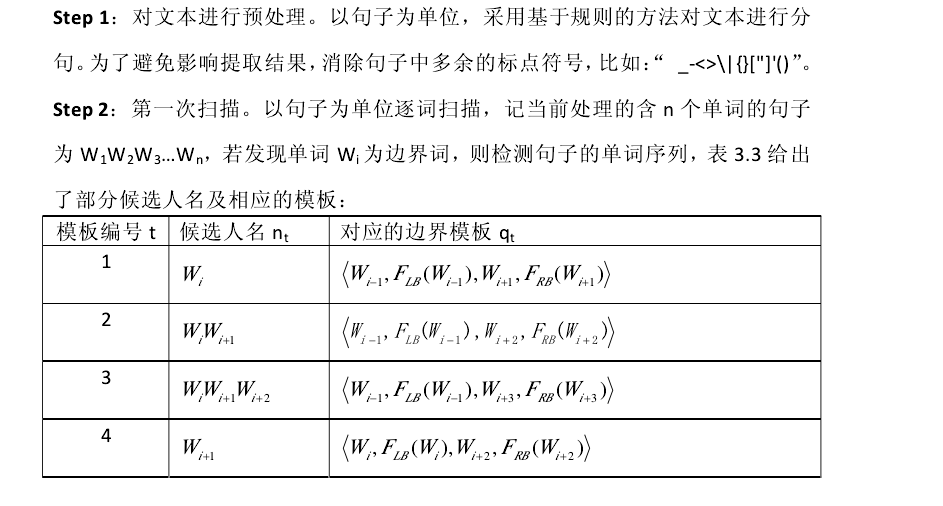
了 1281 个左边界、1364 个右边界，并统计了每个边界的出现频度用以构造边界模板。语料中维吾尔人名和左右边界的结构大概三种：

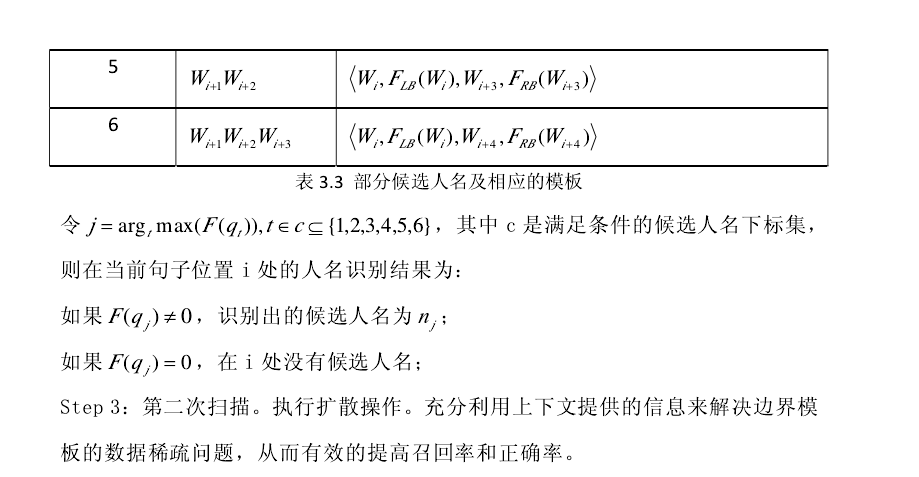
第一种结构：左边界+人名；

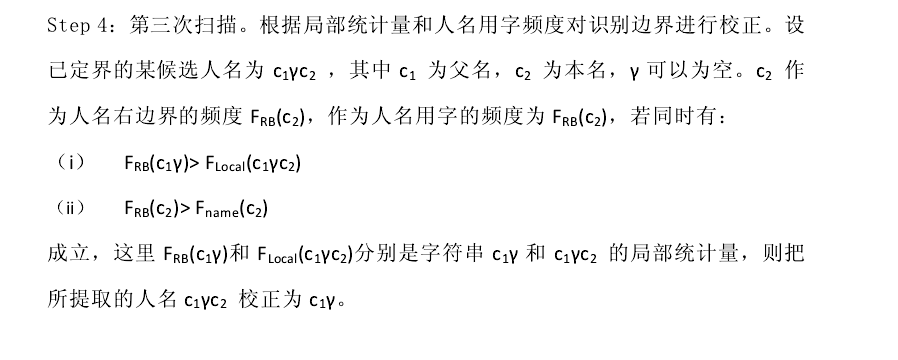
第二种结构：人名+右边界；

第三种结构：左边界+人名+右边界；

候选人名的提取流程：







特征提取和规则施加

跟其它语言的命名实体像似的是，维吾尔人名在文本中不会单独出现，它是

跟其它语言因素一起出现在上下文，而且人名总是依存具有规律信息的上下文。

人名不仅带着一定规律性的信息出现在文本中，而且带着的规律性有时候特别

强。我们通过分析这些规律性信息，总结了维吾尔人名的各种特征，并提取了相

应的识别规则。特征提取完后跟识别规则合理的施加，提高了系统的识别性能。

维吾尔人名的特征

假设，可能作为候选人名的单词为 PWC，它是通过候选人名提取过程提取出

来的，并且 PWC∈context，其中 context 为包含 PWC 的文本。根据维吾尔人名

自身的特点，PWC 的内部具有一定的特征，归纳如下：

（1）人名词典特征

该特征是 PWC 和著名人名词典 FamousPersonNames 中词之间关系的描述，

该特征表示为 PWC∈FamousPersonNames；

（2）普通人名特征

该特征是指：对于 PWC ∉FamousPersonNames 时，PWC∈WomanNames 或者

PWC ∈ManNames；

（3）词缀特征

当 PWC[i...n]∈ManSuffix 时 PWC∈ManNames，或者 PWC[i...n] ∈

WomanSuffix 时 PWC∈WomanNames，其中 PWC[i...n]为候选人名的词缀字符串；

（4）长度特征

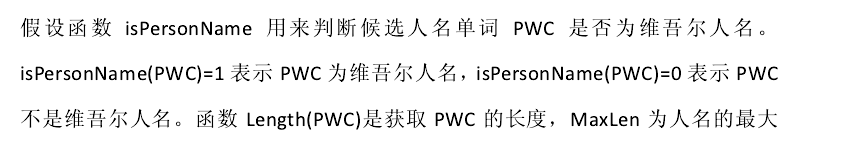
对常用的维吾尔人名库进行统计，结果表明，最长的人名长度是 20 个

字符，最短的人名长度是 3 个字符。

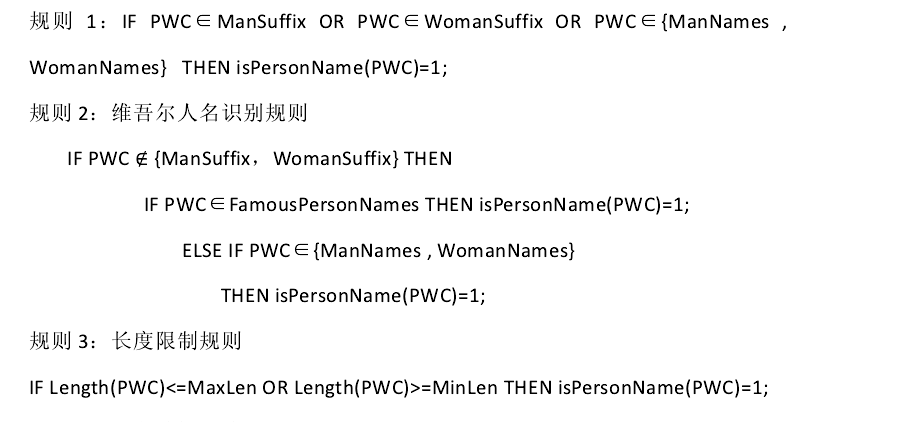
识别规则

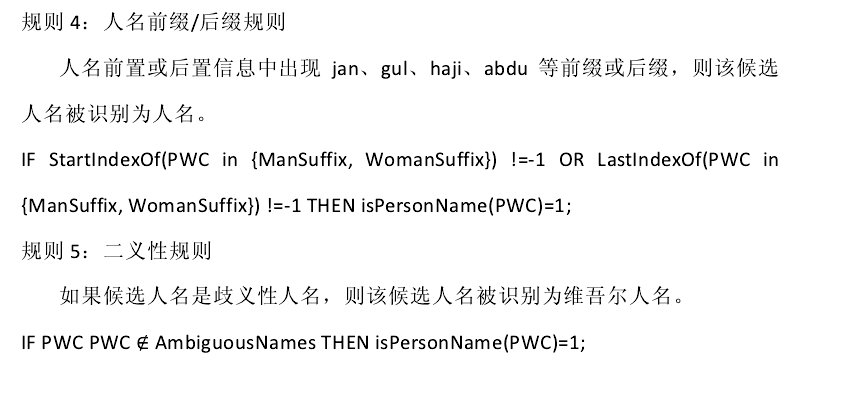
根据上述的维吾尔人名特征，再结合具体实际，本文采用了适合识别维吾尔

人名的 12 条识别规则，其中比较重要的是如下：









消除歧义

消除歧义的规则：

按照上述的词法和语法特性，我们总结了消除歧义的以下几个规则：

1. 并列词规则
2. 称谓词规则
3. 后缀词规则
4. 人名复数规则
5. 句子成分规则
6. 人称规则
7. 否定规则

一些歧义性人名后面加‘la / le / lan’就变成动词，这时无论出现在任何地方都不是人名。

1. 地名/机构名规则。

歧义性人名在地名或机构名特征词前面出现，则淘汰该歧义性人名。

系统模型

此系统的识别步骤如下：

1.  文本预处理。首先对输入的文本进行分句，然后再进行分词。分词后的单

词数组是作为此系统的输入数据，并作为候选人名提取的纯粹数据。

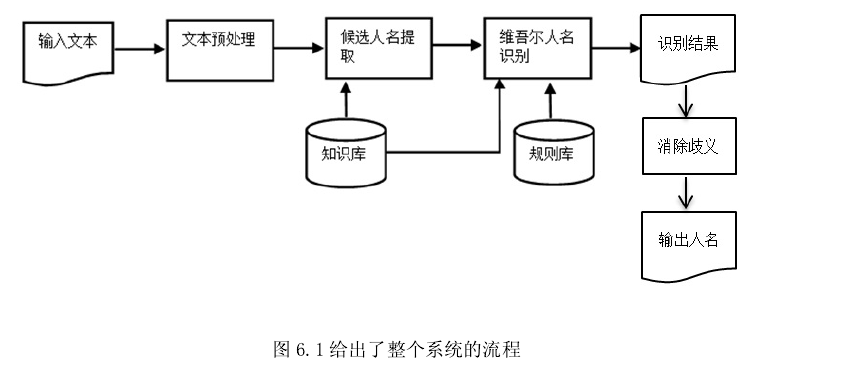
2.  提取候选人名。用提前准备的知识库和统计的方法提取候选人名，并给

下一步的识别提供数据。

3.  规则施加。利用规则集对第二步提取的候选人名进行筛选。

4.  消除歧义。对已识别出的人名进行消歧，进一步提高识别效率和正确性。

图 6.1 给出了整个系统的流程。



实验与结果分析

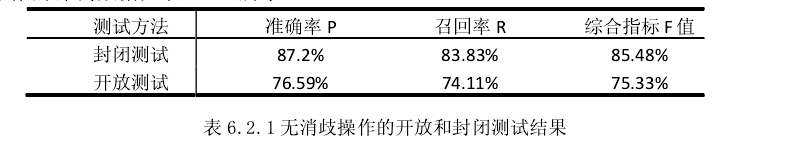
在真实的语料库环境下进行了开放和封闭测试。为测试系统的准确率和

召回率，从“人民日报”和“天山网”采集维吾尔人名较多的 1453（9.41MB）

篇新闻作为封闭测试语料，从“新疆维吾尔作家协会”、“维吾尔作家论坛”等网

站下载维吾尔人名较多的 500（3.18MB）篇文章作为开放测试语料。不进行消歧

操作的评测数据如表 6.2.1 所示：



我们用同样的预料做了消歧实验，评测数据如表 6.2.2 所示：

