汉维时间数字和量词的识别与翻译研究

阿依古丽·哈力克[1,2]，艾山·吾买尔[1,2]，
吐尔根·伊布拉音[1,2]，卡哈尔江·阿比的热西提[1,2]，买合木提·买买提[1,2]

（1．新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046；
2．新疆多语种信息技术重点实验室，新疆 乌鲁木齐 830046）

摘 要:统计机器翻译对时间、数字、量词的泛化能力较弱，为了提高汉维机器翻译系统对时间、数字和量词短语的翻译性能，本文利用双语语料库挖掘并提取汉语时间、数字、量词表达与翻译模式，实现了基于模板的时间、数字、无歧义量词翻译方法及基于上下文的有歧义量词翻译方法。时间、数字、无歧义量词、有歧义量词的翻译 F 值达到了 93.23%、90.15%、96.55%、87.58%，实验证明，本文提出的方法具有简单高效的优点。

关键词：时间数字；无歧义量词；有歧义量词；翻译规则；翻译模板

中图分类号：TP391　**文献标识码：**A

Research on Recognition and translation of Chinese -Uyghur Time and Numeral and Quantifier

Ayiguli HALIKE[1,2]，Hasan WUMAIER[1,2]，Kahaerjiang ABIDEREXITI[1,2]，Maihemuti MAIMAITI[1,2]，Tuergen YIBULAYIN[1,2]

（1．School of Information Science and Engineering，Xinjiang University，Urumqi，Xinjiang 830046，China；
2．Xinjiang Laboratory of Multi-Language Information Technology，Urumqi，Xinjiang，830046，China）

Abstract: The Chinese-Uyghur statistical machine translation system for times, numerals and quantifiers generalization ability are relatively weak. In order to improve the recognition and translation of Chinese-Uyghur machine translation system for times, numerals and quantifiers. This paper uses a method based on corpus to mine and extract of Chinese times, numerals and unambiguous quantifiers, based on context ambiguous quantifiers translation rules .Experimental results show that the translation of times , numerals ,unambiguous quantifiers and ambiguous quantifiers in F value of 93.23%, 90.15%,96.55%,87.58%. The method has advantages of simple and efficient.

Key words: Times, numerals and quantifiers; unambiguous quantifiers; ambiguous quantifiers; translation rules

1 引言

命名实体识别与翻译在机器翻译、信息处理系统中具有重大意义。目前，对命名实体的识别与翻译研究工作国内外已取得了大量的研究成果[1]。Shruti Mathur[2]等用基于规则的方法对英-印命名实体内的常见形式进行了识别与翻译。Deepti Bhalla[3]等使用基于统计的方法识别命名实体，通过平行语料库实现了英语-印语命名实体的翻译。Sameer R. Maskey[4]等通过规则，分析句法开发了英语-阿拉伯语命名实体的翻译系统。Sebastian M P[5]等基于统计实现了英语-马拉雅拉姆语的机器翻译。Feng D[6]等研究了英汉命名实体对齐的新方法。Strötgen J 等基于规则的时间表达式识别与规范化实现了 HeideTime 系统[7]。尹存燕[8]等对中英命名实体识别及对齐研究提出了中文分词优化方法。郑宏[9]、翟飞飞[10]等分别对时间、数字命名实体及数量词的识别问题使用了基于 CYK++算法和语料库、基于规则的方法。李风环[11]等基于面向主题事件构建了时间识别模型。赵紫玉[12-13]分别利用基于短语的翻译模型、规则与 CRF 统计相结合的方法研究日语时间表达式识别与日汉翻译。王伟、李君婵、邬桐等对中文时间表达式的识别与翻译[14-16]提出了基于有限状态自动机，规则库、最大熵，正则文法的方法。杨萍[17]等对汉语-新蒙古文命名实体进行翻译，首先对汉语命名实体进行标注，然后获取汉语 HMM 词对齐结果，最后通过滑动窗口方法选取置信度最高的命名实体翻译。王斯日古楞等对汉-蒙机器翻译系统中的量词短语进行了研究

* **收稿日期：**2016 年 5 月 31 日　**定稿日期：**2016 年 7 月 20 日

基金项目：新疆多语种信息技术实验室开放课题（2016D03023）;国家重点基础研究发展(973)计划(2014CB340506);国家自然科学基金(61331011，61262060，61262061，61063026，61462083)。

作者简介：阿依古丽•哈力克（1991—），女，硕士研究生，主要研究领域为机器翻译、自然语言处理；吐尔根•伊布拉音 (1958—)，男，教授，主要研究领域为机器翻译、自然语言处理；
通讯作者：艾山•吾买尔(1981—)，男，副教授，主要研究领域为自然语言处理与机器翻译；

[18]。邹乐琳等基于统计的方法实现了维语时间表达式的识别[19]。张磊等对维语数词类命名实体（时间、日期、百分比、货币）进行了研究[20]。目前，中国民族语文翻译局在线翻译系统（http://www.mzywfy.org.cn/）和新疆多语种信息技术重点实验室基于短语的 Tilmach 汉维-维汉统计机器翻译系统（http://www.tilmach.cn/）对时间、数字、量词的译文质量不理想。可见，对汉维时间数字和量词的识别与翻译研究工作较少，尤其是汉维方向翻译工作没有针对性的研究。

本文对汉语时间、数字、量词分类及维吾尔文翻译进行详细分析，并构建了相应的时间日期识别与翻译模板、数字识别与翻译模板、无歧义量词词典、基于上下文的有歧义量词翻译规则库，实现了汉维方向的时间、数字、量词识别与翻译算法。

2 汉语时间、数字与量词维吾尔语翻译

汉语是典型的孤立语，属于汉藏语系汉语语族；而维吾尔语是黏着语，属于阿尔泰语系突厥语族，在中国境内使用的维吾尔语是以阿拉伯字母为基础的老维文。汉语维吾尔语语法信息、翻译规律不同，因此对时间数字和量词分别进行讨论。本文研究的老维文在 word 里显示时存在因未安装维吾尔语输入法而导致排版格式混乱的情况，因此维语例子使用拉丁文表示。

2.1 汉维时间日期的分析

由于汉语维吾尔语在各个方面的不同，在汉语-维吾尔语机器翻译系统中，时间表达式的翻译准确率比较低，几乎没有对应的翻译结果。对一些时间表达式的翻译情况对比，如下表 1 所示。

表 1 机器翻译系统的时间日期翻译对比表

汉语时间表达式	Tilmach	中国民族语文翻译局在线翻译系统	维吾尔语正确译文
2007 年 3 月 3 日早上 8 点	<2007 年 3 月 3 日> etigen sa'et 8 de	<2007 年 3 月 3 日> etigen sa'et 8 de	2007-yili3-ayning3-küni etigen sa'et 8de
2016 年 3 月 23 日	2016 年 3 月 23 日	<2016 年 3 月 23 日>	2016-yili3-ayning23-küni
从 7 月到 9 月	7-aydin 9-	<从 7 月到 9 月>	7-aydin9-aygiche

可见，翻译结果出现数字次序混乱、标点符号、词尾丢失或多加等错误情况。原因是时间触发词（年、月、日）在不同的维语时间表达式中译文不同，比如：汉语中“月”对应维语的“yA”，但在不同的时间表达式中翻译结果如下表 2 所示。

表 2 时间触发词的歧义情况表

源语言时间表达式格式	目标语言	源语言	目标语言
6 月	6-ayda	月	Ayda
2013 年 12 月至 2014 年 6 月	2013-yili 12-aydin 2014-yili 6-ayghiche		Aydin
2014 年 4 月 16 日	2014-yili 4-ayning 16-küni		Ayning
6月底	6-aylarning axiri		Aylarning

本文为了解决此问题将时间日期分五类研究，如下表 3 所示。

表 3 汉语-维吾尔语时间日期示例表

时间日期分类	汉语	维吾尔语
时间	明天早上九点半	Ete etigen sa'et Toqquz yärimda
日期	从 2016 年 6 月 4 日开始，截止 5 月 7 日到 9 日	2016-yili 6-ayning 4-künidin bashlap, 5-ayning 7-künidin 9-künigiche

星期，月份，季节	星期四，夏天	Peyshenbe,yaz pesli
周年、年代	60周年	Muqeddes 60 yil
重大纪念日，节日	“古尔邦节”，“肉孜节”，“五·一”国际劳动节	Qurban hëyit Roza hëyit 5-ayning 1-küni Emgekchiler Bayrim

可见，“明天早上九点半”虽然由“明天”-“Ete”，“早上”-“Etigen”，“九点半”-“Toqquz yërim”等三个时间基本单元组成，但在翻译时不能利用时间基本单元组合的方法，应考虑整个模块，在后面加“Da”词尾。“截止 5 月 7 日到 l9 日”同样不能利用基本单元组合的方法翻译，也是应考虑整个模块,翻译结果为：“5-ayning7-künidin19-künigiche”,因为“截止”，“5 月 7 日”，“到”，“l9 日”分别对应为“Giche”和“5-ayning 7-küni”和“Din”和“19-küni”，利用基本单元组合的方法翻译会出现位置不稳定，导致语法错误。

2.2 汉语数字维吾尔语翻译

本文把汉维数字的对比关系分为三种：1）汉语中数字可分为基数词和序数词[21]。维吾尔语中数字分为约数词、集合数词、分数词、序数词、基数词。2）汉语的数字写法有多种形式，比如：对于阿拉伯数字“4”汉语中有“四”、“肆”等写法。维吾尔语有“Töt”一种写法。3）汉语中十位数字的表示形式是“一到九之间的任何一个数”字后面加上“十”而形成的，比如：“二十、三十、四十…”。维吾尔语中这些数字都有专称，比如：“Yigirme、ottuz、qiriq、ellik、atmish、yetmish、seksen、toqsan”。

数字根据翻译结果的不同，总结如下表 4 所示。

表 4 汉维数字翻译示例表

搭配格式	数词类	汉语	维吾尔语译文
数字	纯数字	一,二,一万	Bir,ikki, on ming
前缀+数字	序数词	第一，第二，老二	Birinchi,ikkinchi ,Ikkinchi
数字+量词+后缀	约数词	三十岁左右	Ottuz yashlar chamisida
数字+数字+量词+名词	集合数词	五六个人	Besh -alteylen
数字+分之+数字	分数	五分之一	Beshdin bir
百分之+数字 数字%	百分数	百分之九十	Toqsan Pirsent

可见，汉语的序数词根据意义和计算方法在维吾尔语里分为约数词、集合数词、分数词、序数词，基数词在两个语言中一一对应。根据分类构建规则，对不同搭配格式利用平行语料和翻译模板的方法实现翻译。

2.3 汉语-维吾尔语量词对应关系的分析

量词是表示事物和动作计算单位的词汇。在文献[21]中汉语量词的特点可归纳为如下几种：1）在汉语中根据物体形状的不同，使用的量词也有所不同，因此汉语中的量词较多，大概有 507 个量词。2）在结构上，汉语的量词位于数字和名词之间，数字需要结合量词才能修饰名词。3）根据表示对象的不同，汉语量词分为名量词、动量词、复合量词等三大类。句子中的名量词和动量词不能省略，否则出现语法错误。句子中复合量词（名量词+动量词）中动量词省略掉后不会出现语法错误，但两个句子的意思完全不同。

汉维量词异同点：1）维吾尔语量词没有汉语量词数量多，常用的有“Tal、dane”。2）维吾尔语中量词不能单独做句子的重要成分，它只位于名词或动词的前面，直接修饰名词或动词。3）汉语中有些量词在维吾尔语中没有对应的翻译，即丢失。

在 Tilmach 和中国民族语文翻译局在线翻译系统出现错误情况。例如：量词短语“一线希望”、“一份情”的正确翻译结果应为：“ümid（希望）”、“muhebbet（情）”，即数字和量

词均丢失，但在 Tilmach 中翻译结果是“Azraq ümid”、“Bir parche muhebbet baghlash”，在中国民族语文翻译局中翻译结果是“Azraq ümid”、“Bir ülüsh mihir”。有些事物或行为就必须使用“数字+量词+名词”来表示，量词决不能省略，例如：“十斤面”要翻译为“On（十）jing（斤）un（面）”,不能省略“jing”。汉语根据量词的搭配格式可分为四种，具体如下表 5 所示。

表 5 汉维量词示例表

量词短语格式	汉语量词短语	维吾尔语译文
数字+量词+名词	一群人 一所学校 一峰骆驼 一枚戒指 五条鱼 一个苹果 一本书	Bir top adem Bir mektep Bir tal töge Bir tal üzük Besh tal Bëliq Bir tal alma Bir parche kitab
数字+名词	两兄弟	Aka-ini ikkeylen
量词+名词	双人车	Qosh kishlik mashina
数字+名词+名词	一车瓷砖	Bir mashina sapal xish

可见，量词短语“一群人”、“一所学校”翻译结果都不同，“一群人”的数字、量词、名词全部翻译；“一所学校”翻译数字和名词，量词丢失。“一峰骆驼”、“一枚戒指”…中“峰，枚…”等量词对应的维吾尔语翻译结果为只有一种“Tal”。“一车瓷砖”中“车”是名词，但在量词短语中看成量词。所以，汉语中量词根据翻译需求的不同分为有歧义量词和无歧义量词两大类，有歧义量词指数字一一对应、但量词是一个对应多个（其中包含量词丢失的情况）；无歧义量词指数字一一对应、量词也一一对应，数词一一对应、但量词丢失，数字一一对应、但量词多个对应一个，数字、量词均丢失四种情况。

维吾尔语里面有歧义量词根据句子上下文的不同翻译结果也不同，一个量词有多种翻译结果，如下表 6 所示。

表 6 汉维量词特殊情况示例表

汉语量词	汉语	维吾尔语
头	一头牛	Bir tuyaq kala
	一头狮子	Bir shir
	一头大蒜	Bir bash Samsaq
盘	一盘菜	Bir texse sey
	一盘录音带	Bir daneün'alghulintisi
	一盘电线	Bir yögime tok simi

可见，汉语中量词根据后面的不同名词有不同的翻译结果。“一头牛”、“一头狮子”、“一头大蒜”的“一头”翻译为“Bir tuyaq”、“Bir”、“Bir bash”三种不同结果。同样，“一盘菜”、“一盘录音带”、“一盘电线”的“一盘”翻译为“Bir texse”、“Bir dane”、“Bir yögime”三种不同结果。因此在量词识别与翻译过程中，详细分类会提高机器翻译的准确率。

2.3.1 度量单位

汉维度量单位由数字和量词组成，不需要名词。翻译示例如下表 7 所示。

表 7 汉维度量单位示例表

单位类型	单位名称	维吾尔语译文
长度	公斤，吨，克	Kilogiram,tonna , giram
面积	立方米，立方厘米，ML,L	Kub mëtir,kub santimëtir,ML,L
体积	平方米，立方厘米,CM	Kuwadirat mëtir ,kuwadirat santimitir
重量	分米，厘米，米，公里	Dëstimëtir ,santimëtir,mëtir ,kilomitir
货币	美元，日元，	Dollar,Yin ,yüen ,mo ,koyc hen

	元，角，块钱	

可见，度量单位的翻译为固定译文。

3 汉维时间数字和量词的识别与翻译方法

根据上述汉维翻译规律和语法特点，利用双语语料库挖掘包含时间数字和量词的句子，对每一类分别构建人工编制规则库用于提取汉语时间数字和量词短语，翻译时分别提出了翻译模板、对无歧义量词平行语料库、基于上下文的有歧义量词翻译规则和规则与统计相结合的方式，本方法提高了翻译准确率和召回率。

3.1 汉维时间的识别与翻译方法

根据实际需求对时间表达式没有利用分词系统，而是构建人工编制规则库。即对这些包含前后介词的表达式翻译为维语时，根据不同的时间表达式类型，分别建立了一一对应的 272 个规则和翻译模板。这个方法虽然繁琐，但准确率高，对新闻中时间表达式的覆盖率为 96%。

对时间表达式构造人工编制规则库时，由于维吾尔语和汉语的书写方向相反，为了避免语法错误，把两种语言的规则库分开存储在两个文本文档中。该方法使机器翻译准确率有明显提高。规则和模板一一对应的示例如下表 8、9 所示。

表 8 汉语时间表达式规则库示例表

汉语规则库
(截至\d{1,4}年\d{1,2}月\d{1,2}日\d{1,2}时\d{1,2}分)
(\d{1,4}年\d{1,2}月\d{1,2}日至\d{1,4}年\d{1,2}月\d{1,2}日)
(\d{1,4}年\d{1,2}月至\d{1,4}年\d{1,2}月期间)
(\d{1,4}年\d{1,2}月\d{1,2}日凌晨\d{1,2}时\d{1,2}分)

表 9 维语时间表达式翻译模板示例表

维吾尔语模板
a-yil b-ayning c-küni sa'et d din e minut ötkiche
A-yili b-ayning c-künidin d-yili e-ayning f -küni
a-yili b-aydin c-yili d-ayighiche
a-yil b-ayning c-küni seher sa'et ddin eminut ötkende

可见，汉语规则库和维语模板一一对应。根据上图的人工编制规则识别时间表达式，利用翻译模板抽取对应的翻译结果。过程如下：

输入句子：2013 年 12 月至 2014 年 6 月期间，被告人韦海（广西籍）与境外人员“阿乐”等共谋组织中国境内人员偷渡至越南。

匹配规则后的识别结果：2013 年 12 月至 2014 年 6 月期间。

Tilmach 的译文：<2013 年 12 月至 2014 年 6 月> mezgilide。中国民族语文翻译局的翻译结果为：<2013 年 12 月至 2014 年 6 月> mezgilide。

本方法译文用拉丁文表示：2013-yil 12-aydin 2014-yil 6-ayghiche。与正确译文相同。

对时间日期的识别与翻译完全不依赖各种中文分词系统、标注和统计方法，而是利用人工编制规则库和翻译模板进行识别与翻译，使 Tilmach 的翻译准确率有明显提高。时间日期的识别与翻译处理算法如下图 1 所示。

```
Input: Chinese text set C;
Output :Uyghur Time Expression;
1:  FOR each Ci in C
2:      FOR each Ci matched with Chinese rules:R
3:          IF   Ci∈R   do
4:              Extract the Chinese Time Expression CTi numbers in front of time trigger words a,b,c
5:                  replace To Uyghur Time Expression UTi;
6:          ELSE
7:              CTi
8:          END IF
9:      END FOR
10: END FOR
```

图 1 时间日期识别与翻译处理算法

3.2 汉维数字的识别与翻译方法

Tilmach 对基数词已有了较好的翻译准确率，因此本文解决的是序数词、约数词、集合数词、分数词、百分数、倍数，并对每一类分别构建人工编制规则库，如下表 10 所示。

表 10 数字规则与模板

数字类型	例子	汉语规则特点	维维吾尔语模板
序数词	第三，老大，截止 2026 年	第+数字	a Inchi
约数词	二十左右	数字+介词	a Etrapida
集合数词	五六百人	数字+名词	a+Isim
分数词	三分之二	数字+分之+数字	a din b
百分数	百分之 90	百分之+数字	a pirsent
倍数	三倍	数字+倍	a Hesse

可见，序数词、约数词、集合数词、分数词等都是先用规则来识别数字，然后一一匹配对应的模板库。数字识别与翻译处理算法如下图 2 所示：

Input: Chinese text set C;
Output :Uyghur Numeral Expression;
1:FOR each C_i in C
2: FOR each C_i matched with Chinese rules :R
3: IF $C_i \in R$ **do**
4: Extract the Chinese Numeral Expression: CN_i Translate To Uyghur Numeral
5: expression:UNi
6: And then combine Uyghur Translation UN_i
7: ELSE
8: CN_i
9: END IF
10: END FOR
11:END FOR

图 2 数字识别与翻译处理算法

3.3 汉维量词的识别与翻译

在量词短语的机器翻译中，对收集的语料进行统计分析，根据维吾尔语的翻译规律进行分类，构建量词短语的平行语料库并存储在两个文本文档中。分词对比如下表 11 所示。

表 11 识别方法对比表

对比例子	哈工大分词系统	正确分词结果
一个人	一个 人	一 个 人
一线希望	一线 希望	一 线 希望

为使汉维量词短语的识别与翻译达到尽可能高的覆盖率，本文对无歧义量词识别时利用人工编制规则库识别，对有歧义量词使用哈工大分词系统和人工编制规则相结合的方法，翻译时单独使用模板翻译或统计的方法准确率不高，利用翻译模板和 Tilmach 相结合的方法有更高的准确率。

3.3.1 有歧义量词的识别与翻译

有歧义量词（数字一一对应,但量词一个对应多个）具体的识别与翻译过程如下：利用人工编制规则和哈工大分词系统识别量词短语。翻译时根据哈工大的词性标注结果定位句子里面的名词“/n”来决定这个量词的翻译结果，我们利用这个特点，构建（Special quantifier）平行语料库，从平行语料库中抽取对应的翻译结果，然后结合 Tilmach 实现汉维量词短语翻译。汉语中 45 个有歧义量词，量词“把”翻译结果如下表 12 所示。

表 12 有歧义量词翻译示例表

量词	汉语量词短语	本方法翻译结果	正确译文
把	一把米	Bir siqim	Bir siqim
	一把花儿	Bir deste	Bir deste
	一把年龄	空	空

可见，利用本方法对“一把米、一把花儿”翻译时“数字、量词、名词”均翻译出来，翻译结果为“Bir（一）siqim（把）gürüch（米）”、“Bir（一）deste（把）gül（花儿）”，与正确译文相同。但在 Tilmach 中“一把米”翻译结果“Bir baghlam mätir”是错误的。对“一把年龄”“数字+量词+名词”格式的量词短语，翻译结果仅剩“名词”，“数字、量词”都丢失，翻译结果应为“Yash”。

3.3.2 无歧义量词的识别与翻译

无歧义量词具体的识别与翻译过程如下：首先分为四大类（1.数字一一对应，但量词是多个对应一个；2.数字和量词都丢失；3.数字一一对应，但量词丢失；4.数字一一对应，量词也一一对应），然后利用人工编制规则识别量词短语。

本文分析汉语和维吾尔语的异同点，将量词短语分为下述五种情况。量词识别与翻译处理算法如下图 3 所示。

```
Input：Chinese text set Ci;
Output：Uyghur Quantifier Expression;
1：Quantifier segmentation of each sentences
2：FOR each sentences in the text sets to distinguish the ambiguous quantifier or unambiguous quantifier
3：        IF(ambiguous quantifier)
4：            Match rules and extract the number and quantifier
5：            Translate to Uyghur quantifier
6：        ELSE
7：            Part-of-speech(POS);
8：            IF(POS of sentences has label “/m or /q”)
9：                Match rules and extract the number and quantifier in front of /m and /q
10：               Translate to Uyghur quantifier
11：           End IF
12：       End IF
13：End FOR
```

图 3 量词识别与翻译处理算法

分类一：汉语中的“峰、枚、颗、粒、方、管、则、发、盏、床、灶、柄、梃、梭”等 14 个量词在维吾尔语中的翻译结果都是“Tal、dane（两字意思相同仅写法不同）”。对这种情况解决的方法相对简单，从 Many To One 平行语料库中找出它对应的翻译结果即可。

分类二：汉语中包含“轮、手、桩、宗、阵、记、摊、汪、鸿、团、脬”等 11 个量词的汉语短语翻译为维吾尔语时，短语格式“数字+量词+名词”对应到维吾尔语时仅剩下“名词”格式，即汉语数字和量词都对应为空串。例如：“一轮圆月”、“一手好字”中翻译结果为“Tulun（圆）Ay（月）”、“Yaxshi（好）söz（字）”，“一轮圆月”的“一、轮”、“一手好字”的“一、手”都要翻译为空。

分类三：汉语中“幢、座、扇、堵、所、架、艘、本、家、口、孔、尊、爿、桩、宗、件、罗、首、匝、客、挺、垛、孔、杆、眼”等 25 个汉语量词翻译为维语时没有对应的翻译结果，对这种情况的量词翻译结果使用（Many To Null）平行语料库仅输出数字的翻译结果。例如：在 Tilmach 中“一眼井”结果为“Bir köz quduq”，但正确的结果应为“Bir（一）quduq（井）”，“眼”丢失。

分类四：汉语中 260 个量词翻译为维吾尔语时有对应的翻译结果。例如：“一群人、一出戏、一帖药、一剂药、八味药”中的量词翻译结果分别为“top、meydan、chaplaq、quta、xil”。对这种情况从（One To One）平行语料库中抽取对应的数字、量词翻译结果，然后两个翻译结果合并即可。度量单位（长度、面积、体重、重量、货币）也属于无歧义量词，在格式“数字+单位”后面加或不加名词不会影响翻译结果。

4 实验及结果分析

本文实验的时间数字语料来自“新疆人民日报”、“天山网”的最新新闻，从中自动收集 23447 句包含时间数字的句子,随机抽取包含 9769 个时间数字的 5048 句；量词语料来自“北京大学语料库”，从中收集 39000 句包含量词的句子，随机抽取包含 6723 个量词的 4190 句，收集 1540 句包含度量单位（长度，面积，体积，体重，货币等 105 个）的句子，随机抽取 598 句分别做实验。

在测试语料中，对所有句子进行人工标注、分类、使用平行语料库和翻译模板进行翻译。最后与 Tilmach 进行对比实验，证明了本工作的必要性。

4.1 时间数字与量词实验语料及分析

表 13 含时间数字的汉语语料库表

实验数据	句子数	平均句长	时间数字个数
语料集	23447	78.8	44397
测试集	5048	76.5	9769

表 14 含时间数字语料信息表

时间数字分类	提取语料句数	测试语料（含翻译）	测试时间表达式个数
时间表达式	10920	2268	5633
数字	12527	2780	4136
共	23447	5048	9769

表 15 含量词的汉语语料库表

实验数据	句子数	平均句长	量词个数
语料集	39000	58.3	52065
测试集	4190	52.7	6723

表 16 含量词语料信息表

量词分类	维语数字翻译结果关系	维语量词翻译结果关系	提取语料句数	测试语料句数（含翻译）	测试语料量词个数
有歧义	一一对应	一对多	9520	1052	1896
无歧义	一一对应	多对一	8600	860	1362
	丢失	丢失	5520	552	986
	一一对应	丢失	8440	844	1423
	一一对应	一一对应	6920	882	1056
	共		39000	4190	6723

表 17 含度量单位语料信息表

单位类型	提取语料（含翻译）	测试语料（含翻译）
长度	330	133
面积	247	128
体积	198	60
重量	487	149
货币	278	128
共	1540	598

4.2 评测方法

本文评测指标采用三个值：准确率（P）、召回率（R）、F-Score，计算公式如下：

$$准确率(P) = \frac{正确识别或翻译的数词，量词个数}{系统识别或翻译的数词，量词个数}$$

$$召回率(R) = \frac{正确识别或翻译的数词，量词个数}{总共语料}$$

$$\mathrm{F} = \frac{2 * P * R}{P + R}$$

4.3 实验及结果分析

本文对时间数字和量词采用不同的方法进行测试，然后与 Tilmach 的翻译结果进行对比试验。时间表达式的识别与翻译过程如下：

输入三条句子：

1．截至 7 月 25 日,3 个试点地区共受理。

2．截至 2014 年 6 月 25 日申请 2.3 万余人,同比增长了 5 倍多。

3．全疆 8 月份开始依法全面实施统一的普通护照签发管理政策。

翻译结果用拉丁文表示为：

1．-ayning 25 - axirigha Qeder,bolup üch sinaq nuqtisini rayonda 177 ming 457 kishi qobul qilish.

2．2009 - <2014 年 6 月 25 日> 23 ming adem iltimas,bulturqi shu mezgildikidin besh hessidin artuq köpeydi.

3．8-ayda Shinjang boyiche bashlap adettiki omumyüzlük qanun boyiche yolgha qoyup,bir tutash bashqurush siyasitini pasport bergüchi.

可见，在句 1 中，时间表达式“截至 7 月 25 日”翻译结果的“7”出现丢失；在句 2 中，时间表达式“截至 2014 年 6 月 25 日”没能翻译，“5 倍”后面多加了“%”符号；在句 3 中，“8 月份开始”翻译结果添加了词尾“da”，但应该要添加“din baxlap”。使用人工编制规则库识别时间表达式：“截至 7 月 25 日”、“截至 2014 年 6 月 25 日”、“5 倍”、“8 月份开始”，匹配翻译模板输出翻译结果为：“7-ayning 25-künigiche”、“2014-yili 6-ayning 25-künigiche”、“5 Hesse”、“8-aydin bashlap”，与正确的翻译结果相同。

汉维量词短语的识别与翻译过程如下：

输入句子：夕阳的余辉透过霞云，洒在江心，形成一线闪烁的金斑。

识别：在哈工大的分词系统词性标注结果: 夕阳/n 的/u 余辉/n 透过/v 霞云/n ，/wp 洒/v 在/p 江心/n ，/wp 形成/v 一线/n 闪烁/v 的/u 金斑/n 。/wp。其中“一线”是名词，但汉语翻译为维吾尔语时该句中的“线”应是量词。

翻译：Tilmach 翻译结果用拉丁文表示为：Kechki shepeq bek güzel bolidu diki qalduq nur bulut reng shepeq tumanning singip ötüsh meyxana ,sinxana , yëqinlashmay 1 - sëpide shekillendürüsh bilen chaqnisa ala altun

可见，在 Tilmach 中量词短语“一线”的翻译结果是错误的“1-sëpide”，正确结果“一线”应丢失；匹配平行语料时应按照分类方法中数字一一对应、但量词一对多的情况根据后面的名词来翻译量词短语，即“一线”根据后面的名词“金斑”，翻译结果应为“ ”空串。

4.3.1 识别实验及结果分析

对时间数字和量词考虑句子的上下文信息，建立规则库和模板库，对量词详细分类分别建立平行语料库实现翻译，可以达到比较高的翻译准确率。时间数字、量词识别实验结果如下表 18、19、20 所示。

表 18 时间数字识别实验结果表

时间数字	本方法 P（%）	R（%）	F（%）
时间	97.60	95.85	96.71
日期	93.72	92.76	93.23
序数词	93.87	92.8	93.33
约数词	86.89	87.57	87.22
集合数词	85.65	83.34	84.47
倍数	96.50	94.67	95.58

表 19 量词识别实验结果表

量词分类	数字翻译结果	量词翻译结果	本方法 P（%）	R（%）	F（%）
有歧义	一一对应	一对多	86.54	88.65	87.58
无歧义	丢失	丢失	98.65	96.28	95.11
	一一对应	丢失	98.77	96.78	97.76
	一一对应	一一对应	97.66	95.23	96.42
	一一对应	多对一	92.68	91.90	92.28

表 20 度量单位识别实验结果表

单位类型	本方法 P（%）	R（%）	F（%）
长度	90.65	91.78	91.21
面积	94.44	94.98	94.71
体积	91.45	92.23	91.84
重量	91.40	91.23	91.31
货币	83.90	85.81	84.85

4.3.2 翻译实验及结果分析

根据时间数字和量词的解决方法不同，分别做实验。

表 21 时间数字翻译实验结果表

时间数字	本方法 P（%）	Tilmach P（%）	R（%）	F（%）
日期	97.60	33.08	95.85	96.72
时间	93.72	45.54	92.77	93.23
序数词	93.87	94.50	92.80	93.33
约数词	86.89	92.50	87.57	87.22
集合数词	85.65	76.94	83.34	82.51
倍数	86.50	75.23	84.67	85.58

表 22 量词翻译实验结果表

量词分类	数字翻译结果	量词翻译结果	本方法 P（%）	Tilmach P（%）	R（%）	F（%）
有歧义	一一对应	一对多	76.65	70.73	74.68	75.65
无歧义	一一对应	多对一	82.77	51.87	79.45	81.07
	丢失	丢失	83.33	41.18	81.92	82.62
	一一对应	丢失	93.75	75.78	87.56	90.55
	一一对应	一一对应	94.87	70.44	92.88	93.86

可见，对 Tilmach 不能识别与翻译的部分有歧义量词和无歧义量词应使用人工编制规则、对有歧义量词利用规则与统计相结合的方法使译文质量有明显提高。

表 23 基于规则与统计相结合的实验结果表

实验的方法	翻译 F（%）
基于规则与统计相结合的方法	89.23

表 24 单位翻译实验结果表

单位类型	本方法 P（%）	R（%）	F（%）
长度	96.65	95.78	96.21
面积	95.54	94.88	95.21
体积	93.45	94.43	94.54
重量	94.34	95.66	94.99
货币	89.92	90.01	89.96

表 25 基线系统翻译结果对比表

方法	BLUE 值	时间	数字	量词
Tilmach	BLUE	0.1547	0.3428	0.3486
	NIST	0.7969	1.3021	1.3251
中国民族语文翻译局	BLUE	0.1414	0.4168	0.4258
	NIST	0.7659	0.1411	1.4115
本方法	BLUE	0.9969	0.8854	0.7547
	NIST	11.0778	9.0365	8.9654

表 25 给出了本方法与 Tilmach、中国民族语文翻译局在线翻译系统的性能进行比较。本方法详细分析汉维机器翻译中时间、数字、量词的歧义性、差异性和实际需求情况，对每一部分都分别采用不同的方法：对时间、数字要考虑上下文介词并采用不同的翻译模板，对量词最关键是分类并采用不同的识别与翻译处理算法。所以本文最关键的工作是为提高汉-维机器翻译系统中的翻译准确率对时间数字和量词的分类采取不同的算法进行研究。

5 结语

本文分析汉-维时间数字和量词的差异性，时间表达式中触发词（年、月、日）、数字、量词在不同的维语时间数字、量词短语译文也有所不同、利用统计的方法出现数字次序乱、标点符号、词尾丢失或多加等错误情况。该方法根据它们的特点构建人工编制规则库、汉维翻译模板，

对这些包含前后介词的时间数字翻译为维吾尔语时，根据不同的时间表达式类型构建对应的规则库和翻译模板；对有歧义量词和四种无歧义量词构建五种平行语料库并输出翻译结果，对第一种有歧义量词短语利用规则和统计相结合的方法、后四种无歧义量词利用人工编制规则的方法使翻译准确率有明显提高。该方法与 Tilmach、中文民族语文翻译局在线翻译系统相比，在翻译准确率上有巨大的提高。本方法简单，效率高，目标明确。

未来相关的工作可以在其他少数民族语言中时间、数字和量词方面共享，可以帮助提高整体领域的发展。

参考文献

[1] 赵军．命名实体识别、排歧和跨语言关联[J]．中文信息学报，2009，23（2）：3-17.

[2] Mathur S, Saxena V P．Hybrid Approach to English-Hindi Name Entity Transliteration[J]．Eprint Arxiv，2014.

[3] Deepti Bhalla，Nisheeth Joshi，Iti Mathur，et al.Improving the Quality of MT Output using Novel Name Entity Translation Scheme[C]//2013 International Conference on Advances in Computing，Communications and Informatics(ICACCI)．India，1548-1553.

[4] Maskey S R，Cmejrek M，Zhou B，et al．Class-based named entity translation in a speech to speech translation system[C]//Spoken Language Technology Workshop，2008．Sl．2009：253-256.

[5] Sebastian M P, Sheena K K, Kumar G S. Extension Schemes for the Alignment Model of English-Malayalam Statistical Machine Translator[C]// Proceedings of the 2012 International Conference on Advances in Computing and Communications. IEEE Computer Society, 2012:86-89.

[6] Feng D，Lü Y，Zhou M．A new approach for English-Chinese named entity alignment[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP)．Stroudsburg，PA，2004：372-379.

[7] Strötgen J，Gertz M．HeidelTime：High quality rule–based extraction and normalization of temporal expressions[C]//Proceedings of the 5th International Workshop on Semantic Evalution．USA PA，Stroudsbury:Association for Computational Linguistics，2010：321-324.

[8] 尹存燕，黄书剑，戴新宇，等．中英命名实体识别及对齐中的中文分词优化[J]．电子学报，2015，43(8)：1481-1487.

[9] 郑宏．汉英双向时间数字和数量词的识别与翻译技术[D]．哈尔滨工业大学硕士学位论文，2011，6.

[10] 翟飞飞，夏睿，周玉，等．汉英双向时间和数字命名实体的识别与翻译系统[C]//第五届全国机器翻译研讨会论文集．2009：172-179.

[11] 李风环，郑德权，赵铁军．基于浅层语义分析的主题事件的时间识别[J]．山东大学学报，2015，50（11）：74-80.

[12] 赵紫玉，徐金安，张玉洁，等.规则与统计相结合的日语时间表达式识别[J]．中文信息学报，2013，27（6）：192-200.

[13] 赵紫玉，徐金安，张玉洁，等．日语时间表达式识别与日汉翻译研究[J].北京大学学报（自然科学版），2014，50(1)：180-186.

[14] 王伟，赵东岩，苏婷婷．C-TERN:一种基于 CFSA 的军事新闻文本时间信息处理算法[J]．北京大学学报（自然科学版），2014，50（1）：9-16.

[15] 李君婵，谭红叶，王风娥．中文时间表达式及类型识别[J]，计算机科学，2012，39（11A）：191-194（下转第 211 页）.

[16] 邬桐，周雅倩，黄萱菁，等．自动构建时间基元规则库的中文时间表达式识别.中文信息学报[J]，2010，24(4)：3-10.

[17] 杨萍，侯宏旭，蒋玉鹏，等．基于双语对齐的 汉语-新蒙古文命名实体翻译[J]．北京大学学报（自然科学学报），2016，52（1）：148-154.

[18] 王斯日古楞，斯琴图，那顺乌日图，等．汉蒙机器翻译系统中量词翻译[J]．中文信息学报，2010，24（5）：92-95.

[19] 邹乐琳，吐尔根·依布拉音，麦热哈巴·艾力，等.基于词干提取的维吾尔语事件类时间短语识别[J]．计算机工程与设计，2014，35（2）：625-630.

[20] 张磊，杨雅婷，米成刚，等．维吾尔语数词类命名实体的识别与翻译[J]．计算机应用与软件，2015，32（8）：64-67.

[21] 孙德金．汉语语法教程[M]．民族版．北京语言大学出版社．2012，8（1）.

阿依古丽•哈力克（1991—），硕士，主要研究领域为自然语言处理与机器翻译。
E-mail: 1506867752@qq.com

艾山•吾买尔(1981—)，副教授，主要研究领域为自然语言处理与机器翻译。
E-mail: hasan1479@xju.edu.cn

吐尔根•伊布拉音 (1958—)，教授，主要研究领域为自然语言处理、机器翻译、软件工程。
E-mail: turgun@xju.edu.cn