基于模板的汉维药品用法用量翻译方法研究

阿依图尔苏•喀迪尔[1,2]，艾山•吾买尔[1,2]，尔夏提•艾合买提[1]，解倩倩[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：由于汉语药品种类繁多，人工翻译药品说明书对译员专业水平要求较高，翻译成本较高。为了降低翻译成本，提高效率，本文根据药品说明书的特点，对药品用法用量机器翻译方法进行研究，提出了基于模板的用法用量翻译方法。针对药品说明书的用法用量内容进行分析并分类，自动抽取 240 余条中文模板，人工书写相应的维吾尔文模板和双语词典，模板设计思想来源于正则文法。利用基于模板的方法设计并实现汉维药品用法用量机器翻译系统；最后对每一种情况选择 200 条含药品用法用量句子进行 BLUE 值评测，其中包含用法用量全部内容的句子翻译实验结果 BLUE 值是 0.3146。

关键词：基于模板的机器翻译，用法用量，正则文法，维吾尔文模板

中图分类号：TP391 文献标识码：A

Research on Drug Dosage Chinse- UighurTranslation Method Based on Template

Aytursun KADIR[1,2]，Aishan WUMAR[1,2]，irxat AHMAT[1]， XIEQianqian[1,2]

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

Abstract：Because there are many kinds of Chinese medicines, the manual translation of drug instructions requires a higher level of translators' professional level, and the cost of translation is higher. In this paper, in order to reduce the cost of translation, improve the efficiency, according to the characteristics of the drug specifications, the use of drugs in the use of Machine Translation translation method to study, based on the template usage and dosage translation method. For the drug usage and dosage of content for analysis and classification, automatic extraction of the more than 240 Chinese templates and manual writing Uighur template and a bilingual dictionary, the template design idea comes from regular grammar. Design and Implementation Based on template usage and dosage of machine translation system; finally each choose 200 drug usage and dosage of sentences are blue value evaluation, including the usage and dosage of all content in the blue value is 0.3146.

Key words：Template based Machine Translation; usage and dosage; Regular grammar; Uighur template.

1 引言

药品说明书作为药品相关重要信息的载体，是医生开药和患者用药的重要指南，在日常生活中有着非常重要的作用，尤其是用法用量和规格显得尤为重要。2014 年 9 月份，新浪新闻中心的报告指出，我国每年大约有 250 万人因吃错药而损害健康，甚至由此而导致死亡的人数已达 20 万。吃错药的主要原因就是看不懂药品说明书，尤其是对用法用量、规格和适应症等没有一个正确的概念。新疆气候干燥，很多地方的自然条件恶劣、医疗条件差，这些问题的存在导致民众的患病率较高[1]。新疆是少数民族的聚集地区，少数民族的人口占新疆总人口的 59.9%，此外，具有初中以下文化程度的人数占新疆总人口的 61.0%，而这些人掌握的语言主要是维吾尔语，基本上没有听、读、写中文的能力。截至目前，我国的医疗行业的大部分医疗器件说明书及药品说明书采用中文显示，只有少数新疆内药品企业生产的产品添加维吾尔文药品说明书。除此之外，新疆的 2000 家乡镇医疗卫生院中有 1700 家面向维吾尔族群众，但是这些医院的处方和给用药者提供一个翻译质量较好的说明书是当前非常紧迫的任务。

*收稿日期：2016 年 6 月 25 定稿日期：2016 年 8 月 10 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(61262060，)

作者简介：阿依图尔孙•喀迪尔（1990—），女，硕士研究生，主要研究领域为自然语言处理与机器翻译；
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；
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；

到目前为止我国大约有 16 万种药品，在新疆地区常用的药品达到 1100 种，面对众多的药品，靠人工来翻译药品说明[2]虽然翻译病例也都是中文。因此，大部分维吾尔族群众在医院看病、药店买药时经常会遇到不理解所买的药品的名称、用法、剂量、不良反应、过期日期等问题，在这种情况下很有可能会导致吃错药、多服用、服用过期药品等现象发生，甚至可能造成命案，所以译的效果比较理想，但是会花费大量的人力和时间，相比之下，机器翻译尽管翻译的效果相对于人工而言有些差距但是会大大减少翻译成本，因此本文的研究对社会有着十分重要的意义。

2 机器翻译技术简述

机器翻译[3-10]（MachineTranslation，MT），是利用计算机从一种自然语言到另一种自然语言的翻译技术。目前为止，公认的机器翻译方法分为基于大型语料和统计技术的机器翻译 SBMT(StatisticsBased Machine Translation)，基于规则的机器翻译 RBMT(Rule-based Machine Translation)及基于模板的机器翻译 TBMT（TemplateBasedMachineTranslation）。SBMT 是以大型语料为核心，从数学角度，侧重统计建模，而具备良好的数学模型，鲁棒性以及自学能力，不需要任何的语言现象，需要语料的规模较点。RBMT 是由词典、规则库以及各类知识库构成知识源；从语言现象着手，侧重描述语言构成规律，对语言规律具有良好的概括及描述的能力，但对语法规则较严密的文本，描述其规则较难。TBMT 是对具有句型的覆盖率比较广泛的模板库，它利用翻译模板取代翻译实例，这样既减小了实例库的规模，同时也在一定程度上提高了模板匹配的几率，是一种有效的机器翻译方法。此外，此方法的正确率较高。

对药品说明书中用法用量的翻译研究对社会、尤其是广大消费者来说具有重大的意义，关乎他们的切身利益，甚至生命安全。因此对于译文的质量要求较高[11]，要保证翻译信息的绝对正确性。为此，本文首先利用网络爬虫来收集家庭常备药品信息 10000 条，抽取其中的用法用量，其中通用名称相同的药品信息占总数据的一半多，与此同时，同一个品种，同一种制剂规格，不同企业生产，包装规格不一定相同，药品的质量标准都是一致的，适应症（功能主治）和用法用量的规定都是相同的，因此构建 240 余条药品用法用量模板；此模板在 10000 条用法用量的覆盖率是 60.23%。最后采用基于模板的机器翻译方法来对用法用量进行翻译，得到了比较理想的译文。

3 药品用法用量内容的分析

一般情况下，药品说明书是药品必不可少的部分。药品说明书具有以下特点：句法简单、词汇专业性强，便于用户快速精确地理解、包含消费者应知道的药品信息。例如吴太咽炎片包装盒如下图 1 所示，各部分传达的信息不同，且各自的语言特点和翻译方法也不同，这对中药说明书的翻译工作增加了难度[12]。

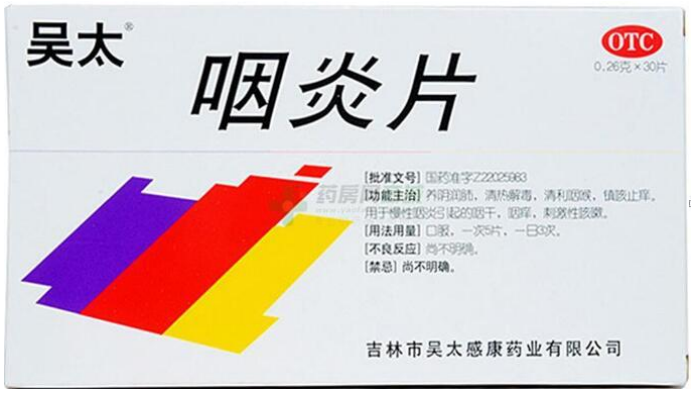

图 1 药品包括的基本内容

在日常生活中，人们通过阅读药品说明书的用法用量来了解该药品由哪些人、什么时候、用什么方式来服用多少量等信息。药品用法用量一般包括四个内容：用药者、用法、剂量、药品服用时间等。其中药品服用适应者有成人、儿童、孕妇和幼儿等。用法一般有口服、含服、注射和滴注等。药物服用时间有空腹、睡前、饭前及其饭后等。用量是指患病者一天内服用的药品次数以及每次的剂量。本文对药品说明书用法用量的以上四个部分进行分析和研究，并提出汉维药品机器翻译方法。在中文待翻译的对象中，有的药品说明书把四个内容分为四个句子，即有单独的服用适应者、用法、剂量、服用时间等，如下表 1 所示。有的说明书一个句子里面包括两个内容，如下表 2 所示；也有的包括三个或者四个，如下表 3 所示

表 1 只包括一个内容的用法用量句子

例子 内容	中文	维吾尔文
用药者	①1 岁以下。 ②成人及 12 岁以上的儿童。	1ياشتىن تۆۋەنلەر① قورامغا يەتكەنلەرۋە 12ياشتىن يۇقىرى باللار②
用法	①口服或开水冲服。	①ئچىلدۇ ياكى قاينىق سۇدا دەملەپ ئچىلدۇ
用量	①一次 100～200mg。	①بىرقېتىمدا100مىللىگرامدىن 200مىللىگرامغىچە
用药时间	①饭前 1 小时服用	①تاماقتىن بىر سائەت بۇرۇن ئىستىمال قىلىندۇ

表 2 包括两个内容的用法用量句子

例子 内容	中文	维吾尔文
用药者+用量	①成人一日 1 片。 ②便秘患者一次 10 克。	قۇرامغا يەتكەنلەر بىر كۈنده بىر تال① قەۋزىيەت بىمارلىرى بىر قېتىمدا 10گىرام②
用药时间+用量	①睡前 1 片。 ②早晚各服 2 粒。	ئۇخلاشتىن بۇرۇن 1تال ① ئەتىگەنده كەچتە 2تالدىن ئچىلدۇ②
用法+用量	①口服，一次 3 粒。 ② 静注 2mg/kg。	ئچىلدۇ،بىر قېتىمدا 3تال① ۋېنادىن ھەر كىلوگرامغا 2مىللىگرام②
用药者+用药时间	①成人一日 3 次。 ②儿童每 8 小时一次。	قورامغا يەتكەنلەر بىر كۈنده 3قېتىم① باللار ھەر 8سائەتته بىر قېتىم ②

表 3 包括三个内容的用法用量句子

例子 内容	中文	维吾尔语
用药者+用量+用法	①成人一日 1.6g（一日 16 片），分 2～4 次服用。 ②成人每 4 至 6 小时服 1-2 袋。	قورامغا يەتكەنلەر بىر كۈنده 1.6گىرام(بىر كۈنىدە① 16 تال)،2قېتىمدىن4قېتىمغا بۆلۈپ ئىستىمال قىلىنىدۇ. ②قورامغا يەتكەنلەر ھەر 4سائەتتىن 6سائەتته 1تالدىن2تالغىچە ئىستىمال قىلىندۇ.
用药者+用药时间+用法	①成人，一日 3 次，饭后服。②儿童：每天 2-3 次，慢速静脉注射。	①قورامغا يەتكەنلەر ،بىر كۈنده 3قېتىم،تاماقتىن كېين ئىستىمال قىلىندۇ. ②باللار،ھەر كۈنى 2قېتىمدىن3قېتىمغىچە،ۋېنادىن ئاستا ئۇرۇلىدۇ.
用法+用量+用药时间	①含服，一次 1 片，一日 4～5 次。 ②开水冲服，一次 1 袋，一日 2～3 次。	①شۇمۇپ ئچىلدۇ،بىر قېتىمدا بىر تال،بىر كۈنده 4 قېتىمدىن 5قېتىمغىچە ئىستىمال قىلىندۇ ②قاينىق سۇدا دەملەپ ئچىلدۇ،بىر قېتىمدا بىر بولاق،بىر كۈنده 2قېتىمدىن 4قېتىمغچە.

4 用法用量的翻译

本章根据上一张所描述的用法用量的内容特征提出翻译方法；此方法中，首先依据用法用量的文本特征构建汉文模板、双语词典及维吾尔文模板，然后进行翻译。

4.1 抽取源语的模板库

药品说明书的用法用量部分一般包括用药者、用法、剂量和用药时间等四个内容。有些说明书包含只有一个内容，有些包含两个或者三个内容。这四个内容中，部分用药者、用量及用药时间的格式固定并且有规律[13]，因此，本文挑选 1195 条药用法用量，构建中文模板，如下表 4 所示。

表 4 中文模板的抽取

例子	抽取
一日 3~4 次	一日[0-9]~[0-9]次
三岁以内一次 6 克	[0-9 0 - 9 一二三四五六七八九十壹贰叁肆伍陆柒捌玖拾]岁以内一次[0-9]克
一次 15g	[0-9 0 - 9 一二三四五六七八九十壹贰叁肆伍陆柒捌玖拾]次[0-9][0-9]g
慢性患者一次 0.5g（7 粒）	慢性患者一次[0-9].[0-9]g（[0-9]粒）

抽取步骤如下：1)输入中文用药品用法用量句子，2)识别提取数字[14]，其他的词作为常量留下，3)编制数字的识别正则表达式，4)选择候选模板，5)手动校正并添加模板库。

4.2 构建双语词典及其译语的模板库

中文药品说明书一般包括上面所说的四个内容：用药者、用法、用药时间、用量（剂量）等。其中部分用药者和用法用词典翻译，比如：成人、儿童、幼儿；口服、汉服、开水冲服等。因此，我们对这些数据进行分析整理并人工构建双语词典。

模板是介于规则和翻译实例之间的一种语言表达形式，也可以理解为一种词汇化的规则。一个模板一般由变量和词汇组成。比如：بىر قېتىمدا aگىرامدىن b گىرامغىچە。人工构建的维吾尔文翻译模板如表 5 所示。

表 5 维吾尔文翻译模板

一次[0-9]～[0-9] mg	مىللىگىرامغىچە bمىللىگىرامدىن a بىرقېتىمدا
分[0-9]至[0-9]次服用	قېتىمغابۆلۈپئىستېمالقىلىنىدۇ bقېتىمدىن a
每[0-9]～[0-9]小时	قېتىم c سائەتتئارلىقىدا bسائەتتىن a ھەر
每[0-9]分钟一次	مىنۇتتابىرقېتىم a ھەر
一次[0-9]～[0-9]袋	بولاققىچە bبولاقتىن a بىرقېتىمدا
每日[0-9]～[0-9]次	قېتىمغىچە bقېتىمدىن a ھەركۈنى

4.3 药品用法用量的翻译方法

图 2 以用法用量翻译为例，给出系统的主要翻译流程。

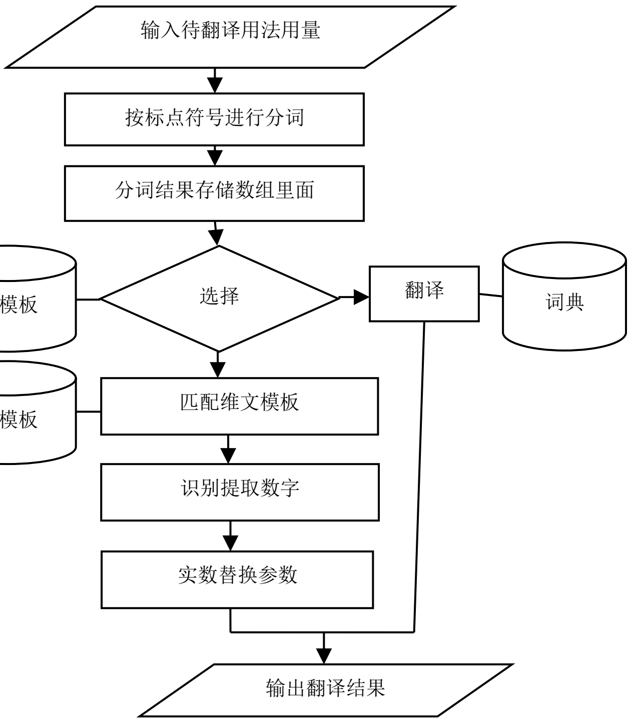

图 2 系统的翻译流程

本系统对一条用法用量句子的翻译流程如下：输入一个用法用量信息“开水冲服，一日 3 次，一次 2~4 袋”；对于这个句子首先按标点符号切分得到“开水冲服”，“一日 3 次”，“一次 2~4 袋”三个短语，然后把这些短语存储在同一个数组里面；对数组的第一个元素“开水冲服”无法利用规则库进行识别，本文使用基于词典的方法进行翻译，数组的第二、第三个元素“一日 3 次”与“一次 2~4 袋”利用基于规则库进行识别，与此同时，从维文模板库里抽取相应的维文模板“ بىر كۈندە aقېتىم ”，“ بىر قېتىمدا aبولاقتىن bبولاققىچە ”，然后自动提取短语“一日 3 次”中的数字 3，替换维文模板里的参数 a，得到译文；再自动提取短语“一次 2~4 袋”中的数字 2 和 4，替换维文模板里的参数 a 和 b，得到译文，最后将三个短语合并后输出翻译结果。三个短语的翻译结果和最终合并后的翻译记过如下表 6 所示：

表 6 最终翻译结果

翻译短语	维吾尔译文
开水冲服	ئچىلدىغان دورا
一日 3 次	بىر كۈندە 3قېتىم
一次 2~4 袋	بىر قېتىمدا 2بولاقتىن 4بولاققىچە
输出译文	ئچىلدىغان دورا،بىر كۈندە 3قېتىم ،بىر قېتىمدا 2بولاقتىن 4بولاققىچە

5 实验及分析

此部分利用三种系统对用法用量进行翻译，比较三个系统的翻译效果。

5.1 实验数据的准备

本文选择了数据结构比较好的网站“药品网”，网站地址如下“http://yao.xywy.com”；收集数据源时，主要是针对“家庭常备”的药品信息抓取。使用网络爬虫收集的药品信息总共有 10000 条家庭常备药品信息，其中消炎药系列 1362 条，感冒系列的药品信息 8381 条，皮肤病系列 257 条。其中抽取 1195 条药品说明书的用法用量中抽取 240 余条模板，人工构建 240 余条识别规则，创建用药者及其用法构成，包含 221 条记录的双语词典。测试语料随机选择 600 条用法用量句子，其中感冒系列的 269 条，消炎系列的 229 条，皮肤病系列的 102 条。

由于表示用法用量的内容一般包括用法，用药者，剂量，用药时间等内容，本文对药品用法用量的四种类型搜集测试语料 600 条，其中一个句子包含一个或两个内容的语料有 200 条，一个句子包含三个内容的语料有 200 条，一个句子包含全部内容的语料有 200 条，并且这些句子进行 BLUE 值评测。

5.2 实验结果及其分析

表 7 基于模板的翻译系统的实验结果

内容	BLUE 值	NIST	句数
用药者	0.4216	1.3121	200 条
用法	0.5326	1.3613	200 条
剂量	0.5246	1.3485	200 条
用药时间	0.5001	1.3124	200 条

表 8 基于模板的翻译系统的实验结果

内容	BLUE 值	NIST	句数
用药者+用量	0.4258	1.4115	200 条
用药时间+用量	0.4119	1.5027	200 条
用法+用量	0.3490	1.4458	200 条
用药者+用药时间	0.3440	1.4476	200 条

表 9 基于模板的翻译系统的实验结果

内容	BLUE 值	NIST	句数
用药者+用量+用法	0.3428	1.3021	200 条
用药者+用药时间+用法	0.4168	1.321	200 条
用法+用量+用药时间	0.3166	1.3458	200 条
包含全部内容	0.3146	1.323	200 条

表 10Tilmach 系统的翻译结果

内容	BLUE 值	NIST	句数
用药者	0.1311	0.6644	200 条
用法	0.3201	0.6745	200 条
剂量	0.1312	0.6578	200 条
用药时间	0.1314	0.6462	200 条

表 11 中国民族译文翻译局翻译系统的翻译结果

内容	BLUE 值	NIST	句子数量
用药者	0.1868	0.9528	200 条
用法	0.3302	0.9527	200 条
剂量	0.1867	0.8925	200 条
用药时间	0.1882	0.9529	200 条

实验结果的错误分析如下：

1）用法的翻译：用法一般只有一种表达形式，比如“口服，含服，开水冲服”；本系统对于用法的翻译实验结果 BLUE 值最高，因为该部分以基于词典的翻译方法为主。

2）剂量的翻译：该部分的数字随着位数的增加，翻译效率降低。比如：把“一次 200-400mg”翻译时，系统提取的数字分别“2、20、200、4、40、400”，使用数字替换模板的参数时，系统识别不出哪个数字是正确的。

3）用药者的翻译：用药者一般有两种表达形式，第一类是“成人，小儿，儿童”，第二类是“1-3 岁，两岁到五岁”。翻译这个部分时，第一类利用基于词典的方法进行翻译，第二类利用基于模板的方法进行翻译。基于模板的方法进行翻译时由于数字的识别不准确使翻译效率不高。

4）用药时间的翻译：用药时间也有两种表达形式，第一类是“饭后，饭前，睡前”，第二类是“睡前 30 分钟，早餐前 1 个小时”。匹配模板时第一类能正确翻译，第二类仅能翻译出来前半部分，后半部分无法识别。由表 8、9 可以看出随着包含内容的增加翻译效率逐渐降低，这是因为随着内容的增多，句子包含的数字的数量增多，匹配模板时先匹配较短的模板，导致遗漏某些词汇的译文。比如翻译“口服，一次 1~2 袋，一日 3 次”句子时，前面的“口服”和“一日 3 次”正确翻译，但是“一次 1~2 袋”翻译这个部分，句子匹配模板时产生错误导致翻译结果不准确，翻译结果如下“ئىچىلىدىغاندورا،2بولاق،بىركۈندە 3قېتىم（口服，一次 2 袋，一日 3 次）”。该例子可以看出，模板匹配时首先匹配最短的“[0-9]袋”模板，匹配不了长的“[0-9]~[0-9]袋”模板而影响翻译结果。

表 10、11 是 Tilmach 和中央民族翻译局翻译系统的实验结果。Tilmach 系统翻译用法用量时产生遗漏错误而导致实验结果降低，比如翻译“早晚各一粒”时，翻译结果是“ئەتگەنۋەبىرتال（早各一粒）”；中央民族翻译局的翻译系统翻译用法用量时同样产生遗漏错误而降低翻译准确率，比如翻译“早晚各一粒”时，翻译结果是“ئەتگەندەۋەبىرتال（早各一粒）”；本系统翻译“早晚各一粒”时，翻译结果是“ئەتگەندەكەچتەبىرتالدىن（早晚各一粒）”。从整个实验结果可以看出，本系统使用的基于模板的机器翻译方法可以提高用法用量的翻译效率，并且此方法补充了 Tilmach 翻译系统和中央民族翻译局的翻译系统的缺点。

6 结语

本文针对说明书的用法用量进行分析并分类，初次实现了基于模板的汉维药品说明书用法用量翻译系统。实验结果表明，本文的方法对药品说明书用法用量的翻译准确率较高，但同时本方法在解决机器翻译问题上还存在以下不足:模板不能递归使用，不够灵活；对于包含数字的用法用量，随着数字的位数增多，模板正确识别数字的准确率降低，导致翻译准确率降低。在下一步的研究中，笔者将从以下几个方面展开工作：1)增加模板的数量，2)准确匹配及识别包含数字的用法用量句子，数字部分使用长度最大的数字 3)改进模板的设计，对用法用量句子结构进一步分析，将模板设计得更加合理，平衡精确度和覆盖率之间的关系，最后，利用基于统计与模板相结合的方法提高译文质量。

参考文献

［1］达瓦・伊德木草，艾山・吾买尔．实例统计翻译混合策略的汉民病历翻译的研究[J].新疆大学学报：自然科学版，2015（1）：68-73．

［2］韦薇，邵觅，林巧彬．我国说明书翻译研究综述[J]．海外英语，2013（3）：7-9．

［3］李业刚，黄河燕，史树敏，等．多策略机器翻译研究综述[J]．中文信息学报，2015，29（2）：1-9．

［4］Nimaiti M，Izumi Y．A rule based approach for Japanese-Uighur machine translation system[C]//Cognitive Informatics& Cognitive Computing（ICCI*CC），2012 IEEE 11th International Conference on．IEEE，2012：124-129．

［5］吴闯．基于模板的汉日机器翻译系统的研究与实现[D]．东北大学，2010．

［6］W.Jing，H.Hou，F.Bao，et al．Template-based model for Mongolian-Chinese machine translation[C]//Conference on Technologies and Applications of Artificial Intelligence．IEEE，2015．

［7］麦热哈巴・艾力．基于实例的维汉机器翻译若干关键问题研究[D]．新疆大学，2014．

［8］Zbib R，Kayser M，Matsoukas S，et al．Methods for integrating rule-based and statistical systems for Arabic to English machine translation[J]．Machine Translation，2012，26（1-2）：67-83．

［9］Sun X，Huang D．Nested Template-based Model for Chinese-Japanese Machine Translation[C]//IEEE，International Conference on Computational Science and Engineering．IEEE，2011：161-166．

［10］李茂西，宗成庆．机器翻译系统融合技术综述[J]．中文信息学报，2010，24（4）:74-84．

［11］杨娅，杨志豪，林鸿飞，等．MBNER：面向生物医学领域的多种实体识别系统[J]．中文信息学报，2016（1）．

［12］刘欢欢．中药说明书中功能用语的汉英翻译方法[J]．科教文汇旬刊，2015（21）：181-182．

［13］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（S3）：191-194．

［14］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．

作者联系方式：阿依图尔孙•喀迪尔,新疆乌鲁木齐市天山区胜利路 14 号,邮编：830046,手机：13579205441，电子邮件：13579205441@163.com

通讯作者联系方式：艾山・吾买尔新疆乌鲁木齐市天山区胜利路 14 号，邮编：830046,手机：13659913514，电子邮件：hasan1479@xju.edu.cn