维吾尔语不同词尾粒度对维汉词对齐的影响研究

麦热哈巴.艾力[1] 米莉万.雪合来提[2] 麦合甫热提[3]
新疆大学 信息科学与工程学院，新疆 乌鲁木齐，830046
E-mail:{marhaba@xju.edu.cn，mihreban@126.com，xmahpu76@163.com}

摘 要：本文我们主要分析维吾尔语词尾在维汉词对齐中充当的角色，通过对维吾尔语词尾采取“分离—丢弃”方案，提高了维汉词对齐以及维汉机器翻译的质量。此方案首先通过词干提取的方法有效地抑制了数据稀疏问题；其次，尽可能地挖掘了维吾尔语词尾所携带的语义信息，使得更多的汉语词得到相应的译文；最后，尽可能地降低维吾尔语词因词干、词尾分离并作为独立对齐单位时出现的句子长度变得过长问题。我们对维吾尔语名词和动词不同范畴的词尾采用此方案，构造了不同的模板，并用到训练集、开发集及测试集，分别做实验。实验结果表明，我们的方法在提高词对齐及机器翻译质量上起到了一定的作用，证明此方法可行并有效。

关键字：词对齐，维汉机器翻译，词尾粒度模板，形态分析

Study on the effect of different granularity on Uyghur Chinese word alignment

Mairehaba Aili, Miliwan Xuehelaiti, Maihefureti
The School of Information Sience and Engineering, Xinjiang University,
Xinjiang Urumqi, 830046
E-mail:{marhaba@xju.edu.cn，mihreban@126.com，xmahpu76@163.com}

Abstract： *We tried to cope with the complex morphology of Uyghur by applying different schemes of morphological word segmentation to refine the word alignment, futher improve the SMT result in Uyghur-Chinese. In this method, we aimed at ,firstly, to minimize the affects of data sparseness by stemming Uyghur words ; Secondly, to produce more refined alignments by aligning Chinese words with Uyghur affixes which has meanings; Lastly, to reduce the over long length of sentence which is the result of regarding affixes as a token of alignment unit. We apply these schemes to the training , developing and test data. Experiment results show, this method plays positive role on improving Uyghur-Chinese word alignment and further machine translation.*

KeyWords： *word alignment，Uyghur-Chinese machine translation, affixes granularity, morphological analysis*

1 引言

一般，词对齐要求句子以“词”为单位分割。因此类似于汉语这样本身不分开写的语言需要先进行分词，而英语等书写时需要留空格的语言就将空格看成是所谓词的分割标记。词对齐的这一要求对于没有形态变化或者形态变化不太复杂的语言来说似乎很合理，也因此得到了准确率比较高的词对齐因而进一步提高了机器翻译的质量。但对于形态变化丰富而复杂的语言，如：阿拉伯语、土耳其语、芬兰语、维吾尔语等，此种方法导致数据稀疏问题。因此，研究者对于有复杂形态的语言采取不同的方法，如：对词进行词干提取后只保留词干[Z.wang,2011]、词干提取并保留词尾[N. Habash,2006]、词干提取并有条件地保留词尾[A. Bisazza, 2009]等，目的就是通过这些方法来降低其数据稀疏带来的影响。

GIZA++是目前最常用的词对齐工具，因与语言无关并开源，深受使用者的青睐。但是GIZA++有着以下问题：

[1] 需要大规模的语料作为训练语料，当语料规模不足时其对齐结果会降低

[2] 对于句子结构不对称的语言对齐效果并不理想

[3] 在形态变化复杂的语言上得到的对齐结果远远不如其他语言[K. Oflazer,2007,2008]

维吾尔语是粘着性语言，其词的构造、句法结构等方面与汉语有着很大的不同，它们属于两个不同的语系。维吾尔语的词具有丰富的形态，形态是在词干之后缀接不同的词尾后生成，且词尾的缀接可以是多层的，从而表示同一个词干的不同语法特性。不仅维吾尔语词干具有一定的语义，其词尾也带有一定的语义信息。维吾尔语句法结构是 SOV 型，而汉语是 SVO 型。

用 GIZA++对齐维吾尔语和汉语词除了受到以上所列的几点约束以外，维吾尔语词尾对词对齐的影响也是需要考虑的问题。因为词尾不仅带来维吾尔语词的不同形态从而导致数据稀疏问题，同时词尾也携带一定的语义信息，传递某种“信息” 。

本文从维吾尔语特有的语言特性出发，以提高维吾尔语与汉语的词对齐准确率为主要目的，分析了维吾尔语形态、不同词粒度对维汉词对齐及维汉机器翻译的影响，同时为了最大限度地发挥词尾对词对齐的积极影响，通过统计及语言特性分析，制定了词尾选择性的保留的方案即“分离—丢弃”方案，来提高维汉词对齐的正确率从而提高维汉机器翻译质量。本文先从分析维汉词对齐存在的问题以及由来出发，统计分析不同词尾在维汉翻译的影响，介绍由此采取的不同词粒度选择方案及构造的模板，最后介绍不同模板对维汉词对齐及维汉机器翻译的影响以及分析。

2 维汉词对齐存在的问题分析

维吾尔语的特性，如：形态变化、语音和谐以及句子结构等都将影响维吾尔语的机器翻译，其中前两个属于词语层面特性，后一个是句法及句子结构层面特性。由于维吾尔语信息处理处于初期研究阶段，句法层面研究才开始，缺乏很多资源和工具，本文主要从维吾尔语的词法层面考虑维汉词对齐问题。

维吾尔语和汉语是属于不同结构的语言，其不同点可以归纳为以下几点：

（1） 维吾尔语的派生功能很强，具有丰富而复杂的形态，通过在词干后缀接词尾可以派生出一个词的不同形态。 如：mektep (学校)、mektipim（我的学校）mektipimde（在我的学校）、mektipimdiki（我学校的）、mektipimdikiler （我学校的人（指同校的人））……可想而知，维吾尔语的形态使得维吾尔语与其他语言之间的对齐，很容易出现数据稀疏。

（2） 维吾尔语遵循语音和谐规律。为此，同一个词尾有不同的变体，当一个词干后接某个词尾时，从词尾变体中根据语音和谐规律选择一个[力提甫.托乎提,2004]。如果将每个词尾看成是一个独立的词尾，显然为形态变化带来的数据稀疏雪上加霜；而如果能将属于统一范畴的词尾统一表示，可以减少一定量的词尾数[麦热哈巴.艾力,2012]。

（3） 维吾尔语的词尾可多层缀接，表示围绕词干原意基础上更广的意思，有时出现一个维吾尔语词可以对应到几个汉语词甚至是一个完整的汉语句子的情况。 如：ölchemlextürelmemsiler? (你们不能进行标准化吗?)一词的结构是 ölchem+lex+tür+el+me+m+siler，它是在词干 ölqem(标准) 后缀接了具有不同语法意义的词尾而形成的。

（4） 汉语是几乎没有形态变化的屈折性语言，其句法结构式为 SVO; 而维吾尔语是典型的粘着性语言，主要的句法结构为 SOV 式。

从以上分析可以看出，对维吾尔语词进行词干提取可以降低数据稀疏的好方法，但只考虑词干不考虑词尾有可能失去词尾所携带的语义。因此[7]中提出对维吾尔语词语进行词干词尾分离，保留词干的同时保留词尾而且对同一范畴的词尾采用统一形式的方法，不仅一定程度上克服了数据稀疏问题，同时利用词尾携带的语义，对齐了更多的汉语词。但是[7]提

出的方法中因为词尾单被看成是一个独立的 token，虽然可以提高对齐准确率以及召回率，但导致句子长度变得过长。有些本身词数较多的句子 token 数量因此变得更多从而在 Giza++ 进行对齐之前被过滤掉。同时，我们注意到虽然维吾尔语的词尾带着一定的语义信息，但是由于维汉语言的不同特性，维吾尔语词尾对应的译文并不是每次都是明文显示，而有时候需要通过上下文或者通过标点符号（表示语调）的形式出现。

这个分析结果让我们进一步研究到底词尾粒度怎样制定才对机器翻译有“更大”的贡献？是不是所有的词尾都需要分离？或者应该选择性的分离？甚至是否丢弃？为了回复这些问题，下面我们从语言学角度及统计角度分析了维吾尔语词尾的一些特性，并根据此特性制定了一系列词尾分离及丢弃的方案，构成不同粒度的词尾并通过实验验证它们对词对齐的不同影响。

3 统计分析维吾尔语词尾在汉语中的译文情况

每种语言都有自己的语法范畴，不能将某个语言的语法范畴强行对应到另一种语言的语法范畴。语法范畴的不同，导致两种语言对同一内容的不同表示方法，进而导致某些语法范畴漏翻、增翻等现象。维吾尔语形态极其丰富，一个词的不同形态不会改变词本身的词汇意义，而给它增加各种语法意义或改变它原有的语法意义。不同形态在维吾尔语中表示不同的语法意义，但汉语中不一定一一体现出来。如下**表**来说明这一点：

表 1 不同维吾尔语句子的汉语翻译

维语句子	含义	翻译
alim het yazdi.	叙述者自己知道这件事	阿里木写信了
alim het yëziptu.	叙述者通过某种手段后来知道这件事	阿里木写信了。
alim het yëziptudek.	叙述者从别处听说有这件事	阿里木好像写信了
u mektepke bardi.	叙述者自己知道这件事	他去了学校。
u mektepke beriwaldi.	表示动作的结果趋向于主体	他去了学校。
u küldi.	叙述者自己知道这件事	他笑了。
u küliwetti.	表示动作是在情不自禁或出乎意料的情况下发生	他笑了
u küliwaldi.	表示动作的结果趋向于主体	他笑了。

表中可以看出，维吾尔语句子中有些词尾虽在本语言中表示某种语法意义，但在汉语中并不一定能准确表达出来，原因主要是维吾尔语中所拥有的范畴与汉语范畴之间的不同。虽然词尾携带者一定的语义信息，但并不是所有的词尾在汉语中都明文翻译，这促使我们为了让词尾对词对齐发挥更好地作用，对词尾进行过滤并找到有用的词尾粒度。为了得到词尾在维汉词对齐中起到的作用，我们对语料做了一系列统计及分析。语料为新疆电视台提供的每日新闻联播为主的新闻语料，维吾尔语小说《故乡》（总 3 册）以及维吾尔语小说《苏醒的大地》（共 2 册）等。新闻联播可体现与当今国内、国外以及百姓生活的点点滴滴，是具有一定的实时性；文学领域一般都被认为能够体现某种语言的表达能力及特色，对于各种词的各种形态以及词尾的出现具有一定的概括能力，这是我们为什么选择这些语料作为统计对象的原因。语料规模及相关数据可参考表 2。

对以上语料进行词法分析后统计了各种词尾的出现频率，**表 3** 为最多出现的 10 个词尾及出现次数，从表格中可以看出，虽然语料不同，但词尾出现的频率相仿。

表 2 不同语料统计信息

语料	领域	句子数	词数	句子长度		
				MaxLen	MinLen	average
新闻联播	新闻	109527	2621371	128	4	23.9
故乡（1）	文学	8762	299774	810	3	34.21
故乡（2）	文学	8996	277603	521	5	30.85
故乡（3）	文学	9215	318900	1249	5	34.6
苏醒的大地（1）	文学	5197	226071	1529	5	43.5
苏醒的大地（2）	文学	5740	236235	487	5	41.15

表 3 不同词尾在语料中出现信息

№	新闻		故乡 1		故乡 2		故乡 3		苏醒的大地 1		苏醒的大地 2	
	词尾	频率	词尾	频率	词尾	频率	词尾	频率	词尾	频率	词尾	频率
1	+i	225875	+i	9330	+i	8101	+i	9735	+i	7154	+i	7218
2	+ni	96141	+ni	5222	+ni	4432	+ni	5046	+nicg	4252	+nicg	4465
3	+nicg	84254	+nicg	4929	+nicg	4262	+lar	4863	+ni	4160	+ni	3901
4	+lar	59423	+lar	4180	+lar	4127	+nicg	4828	+lar	3123	+lar	3340
5	+si	48439	+di	3305	+di	2876	+di	3657	+di	2553	+di	2400
6	+di	44679	+cfa	2707	+cfa	2215	+cfa	2585	+cfa	2425	+cfa	2191
7	+ix	44255	+din	1836	+liri	1853	+din	1965	+din	1855	+din	1884
8	+cfa	33981	+gca	1828	+din	1762	+gca	1895	+da	1488	+gca	1599
9	+din	27870	+si	1788	+gca	1621	+si	1782	+gca	1463	+da	1407
10	+da	26572	+mu	1596	+si	1592	+da	1758	+si	1338	+si	1321

同时，我们将新闻语料维汉句子用 Giza++做了词对齐后，再对词尾被对齐的情况做了统计。统计结果发现，很多词尾虽然在维吾尔语语料中高频出现，但在汉语中并没找到其译文。以下表 4 为词尾中标为有译文及没译文的部分词尾情况：

表 4 不同词尾对齐情况统计

	正确率高的情况		正确率底的情况	
№	词尾	译文及次数	词尾	译文及次数
1	+DA	在 1448，中 856，期间 641，上 503	+i	将 1463，主席 1457，会议 1433
2	+GachGA	由于 6，导致 2，一下子 1	+ning	对 1707，副 1434，发展 923，常委 792
3	+Din	下午 566，上午 499，从 209，向 42	+IHsh	加强 995，的 809，创新 737，变化 700
4	+IHwat	在 112，正在 73，正 43，坚守 8	+IHp	起 555，开始 488，’，’260，加工 170
5	+GA	的 1050，到 701，取得 568，向 426	+IHt	扩大 752，用 398，增强 265
6	+MA	没有 390，没 126，未 70，不同 61，非 24	+IAr	问题 1191，领域 795，人 415，产业 373
7	+sA	就 44，如果 28，了 24，可 13	+IHwal	迎接 32，收到 12， 解救 9，就诊 7
8	+mu	都 159，还 109，虽然 108，吗 53	+IHl	召开 387，举行 255，显示 166
9	+Dim	了 95， 我 68，来 24，感动 16	+IHr	增加 134，回归 111，提高 106，增长 88

10	+la	不仅 355，就 275，只 57	+DHr	交流 643，部署 404，就业 362

通过表 4 中信息以及实际实例的分析发现，有些词尾在语料中的出现次数很高，但不被翻译的概率反而很高。这暗示着，如果运用语言学知识，选择性地选取很可能有译文的词尾而丢弃没有对应汉语译文或很少出现译文的词尾，不仅对挖掘词尾信息、提高对齐的正确率有帮助，而且可以抑制词干—词尾分离而出现过长句子对 GIZA++带来的负面影响。

4 创建不同词尾粒度模版

根据上一节分析，我们对维吾尔语中词尾数最多的两个词类---名词和动词---的词尾采用“分离—丢弃”方案，构成了不同词尾粒度。其中，“分离”是指有译文的词尾与词干分离；“丢弃”是指没译文的词尾需丢弃。不管是“分离”方案，还是“丢弃”方案，都是通过统计分析以及语言特征基础之上进行，没有绝对的对或错之别，我们的目的是尽可能地发挥词尾对词对齐的正面影响的同时通过去掉翻译概率较低的词尾来克服句子长度问题。

维吾尔语词类中名词和动词的词尾数量最多，因此我们就选择了这两个词类的词尾作为研究对象。受篇幅的影响，下面我们就选择性的列出几个方案：

- 名词的数范畴

数范畴表示名词的单数和复数。单数没有词尾，复数词尾为“-lar，-ler”，但汉语中复数除了人名后加“们”外，常常通过上下文来区分，这一点与维吾尔语不同。所以对数范畴采取了“丢弃”方案。如：

Nurghurn（很多） kitab（书）+lar 很多书

Jiq（许多） Alma（苹果）+lar 许多苹果

- 名词的格范畴

维吾尔语的格范畴有 10 个不同类型，其词尾数达到 22 个，分别为：主格、宾格、领属格、时位格、向格、从格、界限格、相似格、量似格、范围特征格等。根据分析，对领属格和宾格采用了“丢弃”方案。原因是虽然维吾尔语中领属格词尾“-ning”对应着汉语中“的”，但汉语中很多情况下“的”以隐含的方式显现，如：

Dölitimiz（我国）+ning（的） bayliqi mol 我国（的）资源丰富

Balilar öy（房子）+ning（的）sirtida oynawatidu. 孩子们在房子（的）外玩。

类似，宾格词尾“-ni”虽然可以对应到汉语中“把”，但维语句子翻译成汉语时，很多情况下并不直接使用“把”字。如：

Pul(钱)+ni tapshurup aldim. 钱我收到了。 （直译： 我把钱收到了）

Balamni yeslige apirip qoydum. 我送孩子去幼儿园了。（直译：我把孩子送到幼儿园了）。

以此类推。其他时位格、向格、从格、界限格、相似格、量似格、范围特征格等格形式分别对应到汉语中的“在、向、往、从、像”等，采用“分离”方案。

- 动词的否定范畴

维吾尔语动词词尾中，属于否定范畴的词尾为“-ma，-me，-may，-mey，-mas，-mes，-masliq，-meslik”等，几乎都可以翻译成汉语的“不，没，没有，未”等表示否定意思的词，所以对其采取了“分离”措施。

- 动词的语态范畴

维吾尔语动词的“语态”属于动词词干后最先出现的语法范畴，表示动作行为与主体或客体之间的各种关系。维吾尔语动词的语态有 5 种，分别为主动态、使动态、反身态、被动态、交互共同态。其中反身态及被动态在汉语译文出现率较低，虽然被动态可以对应到汉语中“被”字，但汉语中对于事物的描述用主动方式比被动方式多得多而且也是一种习惯。如：

Oqughuchilargha kitab tarqitip bërildi (ber+il+di).

直译：学生们被发书了

正确：给同学们发书了。

Ana balilirini kiyindürdi （key +in +dür +di）.

直译：母亲让孩子们穿上了衣裳。

正确翻译：母亲给孩子们穿了衣裳。

如果再把“给同学们发书了”翻译成维语，往往是“Oqughuchilargha kitap tarqitip berdi.”。同理，交互共同态以及反身态所表达的语法信息在汉语中也不是直接体现出来。根据以上分析以及结合统计分析，我们对反身态、被动态、交互共同态采取了“丢弃”方案，对其余的采取了“分离”方案。

根据以上“丢弃—分离”方案，我们构造了不同词尾粒度的模板。为了找到最有效的词尾粒度方案，我们对同一个类的不同模板赋予了序号，序号大者包括其前一模板的方案，同时用字母表示其类型，如：M 表示模板，N 表示名词，V 表示动词，同时模板 MN_2 包含了模板 MN_1 采用的方案意外，又增加了新的方案。最后将名词词尾最后的模板和动词词尾最后的模板合起来作为新的模板。以下为不同模板的标示符及采用的规则：

MN_1：采用了名词“格”范畴的方案；

MN_2：在 MN1 的基础上增加了名词“人称”范畴的方案；

MN_3：在 MN2 的基础上增加了名词“数”范畴的方案；

MV_1：采用了动词“体”范畴方案；

MV_2：在 MV1 的基础上采用了“语态”范畴方案；

MV_3：在 MV2 的基础上采用了“形动词”方案；

MV_4：在 MV3 的基础上采用了“副动词”方案；

MV_5：在 MV4 的基础上采用了“动名词”方案；

M_{ALL}：采用了 MN3 方案和 MV5 方案。

我们将以上方案分别应用到用于词对齐双语语料的维吾尔语端。如，以 M_{ALL} 为例，对句子“Tuqqanlargha bayramliq sowgha sëtiwaldim .（我给亲戚买了生日礼物。）”采用此模板后进行对齐，使用模板前后的对齐结果分别为图 1 和 图 2，可以发现，使用模板后原先只做形态切分的句子不仅长度有所减少而且原先未能一一对齐的词尾差不多都找到了对应的译文。

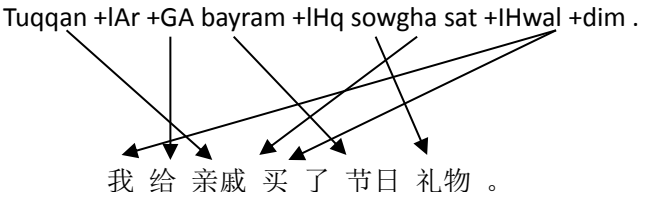

图 1 使用 MALL 模板之前的对齐

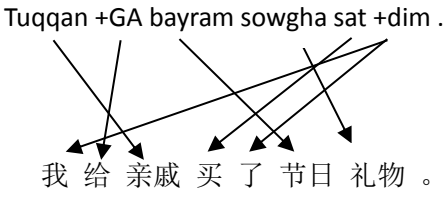

图 2 使用 MALL 模板之后的对齐

5 实验及分析

为了验证“丢弃—分离方案”对维汉词对齐的影响，我们设置了两种实验，实验一着重分析方案对对齐错误率的影响，实验二中分析了此方案对机器翻译起到的作用。

5.1 实验一、“丢弃—分离”方案对对齐错误率的影响

实验中使用了 CWMT2013 提供的维汉新闻领域平行语料，包含 11 万条句对，并使用 GIZA++做词对齐。为了评价词对齐的结果，采用了 AER（对齐错误率）评价标准。

为了发现不同词尾粒度对维汉词对齐的影响并找到最“理想”的词尾粒度，我们将以上 9 中模板依次应用到源语言端，并与目标语言—中文语料构成平行语料，总共构成了 9 对训练语料。为了得到 AER 的结果，还从每种训练语料中随机挑出 100 条句子做手工对齐，作

为标准答案。同时，对维吾尔语句子做词法分析后词干词尾分离并保留所有词尾的情况作为基线，依次计算每一个模板的 AER 值并与其做比较。实验结果为如表 5 所示：

表 5 不同模板 AER 值比较

预处理	模板	*token*	P_r（%）	R_e（%）	*AER*（%）
BaseLine	*BL*	3111214	24.48	28.32	73.73
N_{case}	MN_1	2930844	41.30	52.85	53.63
$N_{case+person}$	MN_2	2679457	44.94	72.13	44.62
$N_{case+person+num}$	MN_3	2618452	57.96	81.17	32.37
V_{aspect}	*MV1*	2862425	38.97	77.22	48.20
$V_{aspect+voice}$	*MV2*	2724058	57.09	80.73	33.11
$V_{aspect+voice+parti}$	*MV3*	2721787	58.46	82.87	31.44
$V_{aspect+voice+part+ad-v}$	*MV4*	2667133	60.04	82.95	30.34
$V_{aspect+voice+part+ad-v+ger}$	MV_5	2666935	61.24	83.12	29.47
V_{all}+N_{all}	M_{ALL}	2402943	63.67	83.57	27.72

分析实验数据，首先注意到语料中标记数（token）的变化。每采用一种模板，token 数都有所下降，说明通过模板的使用丢弃了一些“无用”的词尾，降低了句子长度。同时也可以看出，名词性模板引起的 token 数量下降幅度大于动词性模板。因为任何一种语言的句子中，动词作为中心词出现，而名词、副词、形容词等的出现次数大于动词；AER 的值都是下降的趋势，说明“丢弃-分离”方案对词语对齐起的作用是积极的，AER 的值逐渐变小，只是不同方案对降低 AER 的贡献不同。

5.2 实验二、“丢弃—分离”方案对机器翻译的影响

实验目的是考查方案对机器翻译起到的影响，语料仍然是 CWMT2013 提供的面向新闻领域的维汉训练语料，规模与实验一相等，开发集为 700 条句子，测试语料为 1000 条句子构成。为了分析不同模板对机器翻译的影响，我们按每一种方案重新构造训练语料、开发集及测试集并分别作了翻译。实验中，把新疆多语种信息技术重点实验室参加 CWMT2013 新闻领域维汉机器翻译评测结果作为基线实验，基线实验使用的语料与我们使用的语料相同，但语料中做词法分析后只留下词干并把所有的词尾丢弃。我们用统计机器翻译开放平台 Moses[1]为基线翻译系统，对翻译结果评价标准使用基于词的 BLEU[K. Papineni,2002]值。 系统中，语言模型是利用工具 SRILM[A. Stolcke,2002]训练的五元模型，而训练数据是相应训练集的中文部分，其他参数都没改变，采用默认值。实验结果为**表** 6 所示。

表 6 不同模板对机器翻译的影响

方案	模板	BLEU	幅度
CWMT2013	*BL*	0.4512	基线
N_{case}	MN_1	0.4517	+0.5%
$N_{case+person}$	MN_2	0.4601	+0.89%
$N_{case+person+num}$	MN_3	0.4624	+1.12%
V_{aspect}	*MV1*	0.4407	-1.05%
$V_{aspect+voice}$	*MV2*	0.4515	-1.48%
$V_{aspect+voice+parti}$	*MV3*	0.4602	+0.90%
$V_{aspect+voice+part+ad-v}$	*MV4*	0.4605	+0.93%
$V_{aspect+voice+part+ad-v+ger}$	MV_5	0.4614	+1.02%
V_{all}+N_{all}	M_{ALL}	0.4628	+1.16%

1 http://www.statmt.org/moses/

表中可以看出，不同模板对机器翻译的影响不同。按词性来分，名词性模板的影响比动词性模板大，都是提高 BLEU 值，而且 MN_3 的影响最明显，提高幅度达到了 1.12%；相比而言，动词性模板的影响不是很理想，最好的结果也没达到基线实验值，通过分析，我们认为主要原因为以下：（1）这与不同词性在句子中出现的次数有关，前面已经讨论过，再次不再阐述；（2）动词的形态非常复杂，自动进行词法分析难免携带一定的错误分析的情况，特别是形态结构比较复杂的情况下这种错误就较明显，此错误会蔓延到模板的使用，这也会导致动词影响力。

综上所述，“丢弃—分离”方案，对于形态复杂而词尾携带一定“信息”的维吾尔语而言是可行的，它通过“分离”方案尽可能地保留有意义的 token，同时通过“丢弃”方案将没有意义的词尾丢弃，从而降低句子长度，最终提高机器翻译的质量。但是，通过影响力的分析也可以知道，目前的方案对 BLEU 值的影响虽然是正面的，但幅度不高，这说明此模板的选择有待进一步改善，通过进一步统计并分析后提出更合理的模板也是我们进一步研究的目标。同时，目前我们只考虑了名词和动词，除此之外，还有副词和形容词也是下一步考虑的目标。

参考文献

Z. Wang, Y. Lu, and Q. Liu. 2011. Multi-granularity word alignment and decoding for agglutinative language translation. In *Proceedings of MT SUMMIT*. Pages 360–367.

N. Habash and F. Sadat. 2009. Arabic preprocessing schemes for statistical machine translation. In *Proceedings. of the Human Language Technology Conference of the NAACL*. Pages 49–52.

A. Bisazza and M. Federico.2009. Morphological pre-processing for Turkish to English statistical machine translation. In *Proceedings of workshop on Spoken Language Translation*. Pages 129–135.

K. Oflazer and I. D. El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT*. Pages 25–32.

K. Oflazer. 2008. Statistical machine translation into a morphologically complex language. Computational Linguistics and Intelligent text processing. Pages 376–387.

力提甫.托乎提. 2004. 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, vol. 31, no. 5. Pages 108–113.

麦热哈巴.艾力, 王志洋, 吐尔根.依布拉音.2012. 一种提高维吾尔语_汉语词语对齐的方法研究, 小型微型计算机系统, vol. 33, no. 11. Pages 2551–2554.

K. Papineni, S. Roukos, T. Ward, and W. Zhu.2002. BLEU: a method for automatic evaluation of machine translation. In *Proceeding of ACL* . Pages 311–318.

A. Stolcke. 2004. SRILM-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*. Pages 901–904.

M. Popovic and H. Ney. 2004. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of the 4th International Conference on LREC.*

麦热哈巴.艾力, 姜文斌, 王志洋, 吐尔根.依布拉音,刘群. 2012. 维吾尔语词法分析的有向图模型, 软件学报, vol. 23, no. 12. Pages 3115–3129.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL. Pages 160–167.*

S. Goldwater .2005. Improving Statistical MT through Morphological Analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Languages (HLT/EMNLP)*. Pages 676–683.