语料预处理对蒙古文-汉文统计机器翻译的影响

李金廷 侯宏旭 武 静 王洪彬 樊文婷
(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 传统蒙古文形态分析主要采用将蒙古文词缀和词干直接切分而仅保留词干的方法,该方法会丢掉蒙古文词缀所包含的大量语义信息。蒙古文词缀中包含大量格的附加成分,主要表征句子的结构特征,对其进行切分并不会影响词汇的语义特征,若不进行预处理则会造成严重的数据稀疏问题,从而影响翻译质量。因此,基于现有理论对语料预处理方法进行总结研究,重点研究了蒙古文格处理对翻译结果的影响,目的是从蒙古文形态分析的特殊性入手来提高蒙古文-汉文统计机器翻译的质量。通过优化预处理方法,使机器翻译结果的BLEU得分相比基线系统1提高了3.22个点。

关键词 统计机器翻译,语料预处理,蒙古文形态分析,格处理,拉丁转写,中文分词

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.047

Effect of Preprocessing on Corpus of Mongolian-Chinese Statistical Machine Translation

LI Jin-ting HOU Hong-xu WU Jing WANG Hong-bin FAN Wen-ting
(College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

Abstract The traditional methods of morphology preprocessing use Mongolian suffix segmentation and stemming, which leads to semantic loss of the words. The additional components of Case is a special additional component of the Mongolian word suffix which only represents the syntactic information of the sentence but not the semantic information of the words. Inappropriate preprocessing of the Case causes data sparsity to the machine translation training. Therefore, we summarized and researched the existing corpus preprocessing method of Mongolian morphology to compare the results. Our methods mainly focus on the effect of Case processing and improve the performance of Mongolian-Chinese SMT system of 3.22 relative BLEU score compared to the baseline system.

Keywords Statistical machine translation, Corpus preprocessing, Mongolian morphological analysis, Case processing, Latinization, Chinese word segmentation

自21世纪以来,人工智能领域不断取得新进展,自然语言处理作为人工智能领域的重要研究方向,也取得了较大突破。二十世纪七八十年代学者们开始研究蒙古文-汉文统计机器翻译,并不断提出新的理论和方法,使得蒙古文-汉文翻译的准确率得到不断的提升,其中对语料的预处理是蒙古文-汉文统计机器翻译的基础,也是能否取得高质量翻译结果的关键。世界上许多语言的词存在多种变化形式,词型的变化是一个语法过程,对语义的影响[1]较小。蒙古文是粘着性语言,它的构词和构形都是以词根、词干上连接不同词尾来完成的[2],因此蒙古文词较丰富,形态构成复杂,蒙古文形态分析是预处理的关键。

在蒙古文语料形态分析方面的现有技术主要有词缀的切分以及词干的提取,但蒙古文中存在大量格的附加成分,对其进行处理可以缓解数据稀疏问题,提高翻译质量。本文采用并进行分析的蒙古文预处理方法有:词缀切分、词干提取、格的附加成分的切分和去除以及拉丁转写。在词缀切分方面,对蒙古文切掉词缀之后,采用只提取词干、提取词干并保留一个词缀和提取词干并保留两个词缀3种不同处理方法来进行比较分析b在格的附加成分处理方面,采用切分但保留格的附加成分、切分不保留格的附加成分两种方法来进行比较分析。为了便于处理蒙古文,一般采用拉丁文转写的蒙古文语料。本文采用的转拉丁方法还对蒙古文编码的错误进行了简单的校对;在中文处理方面采用了基于字和基于词两种粒度的切分方式。我们对双语语料分别进行预处理,组合出具有代表性的20组语料,分别使用Moses进行蒙古文-汉文统计机器翻译实验,最终将各组语料的翻译结果进行对比。本文

到稿日期:2016-09-29 返修日期:2016-12-13 本文受国家自然科学基金项目:跨汉斯拉夫蒙古文的信息检索关键技术研究(61362028),内蒙古自治区研究生科研创新项目:蒙古文-汉文语料预处理关键技术的研究(11200-12110201)资助。

李金廷(1994-),男,硕士生,主要研究方向为自然语言处理,E-mail:justin_63@sina.com;**侯宏旭**(1972-),男,博士,教授,主要研究方向为自然语言处理、信息检索,E-mail:cshhx@imu.edu.cn(通信作者);**武 静**(1989-),女,博士生,主要研究方向为自然语言处理,E-mail:wujingyaya@163.com;**王洪彬**(1989-),男,硕士生,主要研究方向为自然语言处理,E-mail:whongbin@mail.imu.edu.cn;**樊文婷**(1992-),女,硕士生,主要研究方向为自然语言处理,E-mail:1583679655@qq.com。

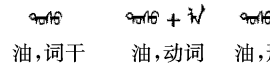
将蒙古文原始语料(未经过任何处理的蒙古文语料)与按词切分的汉文语料作为基线系统 1,将蒙古文原始语料与按字切分的汉文语料作为基线系统 2。将蒙古文进行处理后分别与按词和按字两种不同粒度切分的汉文组成的两组平行双语语料的实验结果分别比基线系统 1 和基线系统 2 的实验结果提高了 0.78 和 0.38 个 BLEU 点;将蒙古文进行格处理再进行拉丁转写之后分别与按词和按字两种不同粒度切分的汉文组成的两组平行双语语料的实验结果分别比基线系统 1 和基线系统 2 提高了 2.5 和 2.04 个 BLEU 点;此外,由于语料规模较小,词缀切分方法中对格处理的不完善等因素导致词干的提取造成了严重的数据稀疏问题。采用蒙古文词干与按词和按字两种不同粒度切分的汉文组成的两组平行双语语料的实验结果比基线系统 1 和基线系统 2 分别提高了 0.4 和 0.17 个 BLEU 点;经过格处理和拉丁转写之后的蒙古文词干语料与按词和按字两种不同粒度切分的汉文组成的两组平行双语语料的实验结果分别比基线系统 1 和基线系统 2 提高了 0.5 和 0.43 个 BLEU 点。其中进行格处理再进行拉丁转写的蒙古文与按字切分的汉文组成的语料的实验结果比基线系统 1 提高了 3.22 个 BLEU 点,是蒙古文汉文统计机器翻译语料预处理最为有效的处理方法。

本文第 1 节主要介绍蒙古文语料方面的形态分析,包括词缀的切分、词干的提取、格处理与拉丁转写;第 2 节将介绍汉文语料预处理的相关工作;第 3 节主要分析对比实验结果,通过实验数据验证所提理论的优越性,并对实验中的问题进行分析和解决;最后总结全文。

1 蒙古文形态分析

1.1 蒙古文词切分

蒙古文是粘着性语言,它的构词和构形均是通过在词根、词干上连接不同词尾来完成的。每个词的构成和其语法意义的表示都依赖于不同词尾的缀接,因此只有正确切分词根、词干和词尾才能揭示其词类属性和语法关系[2]。图 1 所示为中文中"油"在蒙古文中通过连接不同的词缀来表达不同的词性。

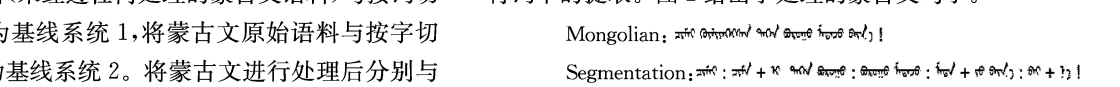


油,词干(原始形式) 油,动词(擦油) 油,形容词(油的)

图 1 蒙古文词干和构词附加成分

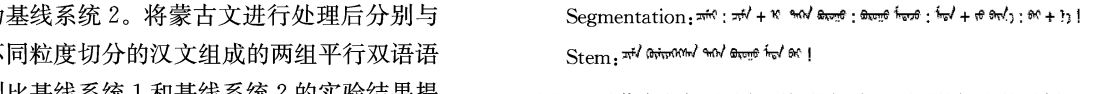
1.2 词干的提取

词干提取是一种将变体词形式映射到其基本形式的过程。它是在语言建模、自然语言处理和信息检索应用程序中使用的基本文本预处理方法之一[3]。蒙古文词形态丰富,其中包含很多变体词,对语料进行词干提取是蒙古文语料预处理的关键技术之一[4]。通过对前人工作的总结(其中包括那顺乌日图等人[2]提出的蒙古文词根、词干、词尾的自动切分;侯宏旭等人[5]提出的基于统计语言模型的蒙古文词切分;赵伟等人[6]提出的基于条件随机场的蒙古文词切分;明玉等人[7]提出的基于词典、规则与统计的词切分等方法)以及对实验结果的分析,最终选择基于词典、规则与统计的词切分的方法对本研究领域的语料进行词切分,之后将切分后的语料进行词干的提取。图 2 给出了处理的蒙古文句子。



Mongolian:

Segmentation:



Stem:

图 2 对蒙古文句子进行词缀的切分和词干的提取的示例

明玉等人提出的方法在词干提取的过程中并未对格的附加成分进行处理。由于统计机器翻译的过程中蒙古文空格被识别为空格,使得格的附加成分与词干分离,单独形成了一个或多个词,从而造成了严重的数据稀疏问题。我们对语料中的词和词干进行频率分析,由于拉丁转写可以纠正一定的语料错误,因此本文均是在蒙古文语料拉丁转写的基础上进行蒙古文语料的词频分析。蒙古文语料库中词和词干的频率分布如表 1 所列。

表 1 蒙古文词和词干的频率分布

	蒙古文词	蒙古文词干
总词数	42781	34335
一次词频总数	23734	20818
二次词频总数	6216	4812
三次词频总数	2840	1988
多次词频总数	9991	6715
一次词频占比/%	55.5	66.6

表 1 中的数据显示,蒙古文词干语料中仅出现一次的词干的频率比蒙古文词语料中仅出现一次的词的频率高 11%,数据证明经过词干提取之后的蒙古文语料存在严重的数据稀疏问题。虽然词干的提取可以使蒙古文-汉文统计机器翻译的过程中的词对齐效果得到提升,但是蒙古文-汉文统计机器翻译的语料规模较小,数据稀疏问题较严重,且词缀的去除势必会造成信息的丢失,因此我们假设在蒙古文-汉文统计机器翻译过程中由词干的提取带来的问题将影响最终的翻译质量。

1.3 格处理

蒙古文控制符有 7 种,本文处理的控制符主要包括图 3 中所列举的 4 种,我们通过编码方式来识别这 4 种控制符,并对其进行频率的统计。其中第 4 种窄宽度无间断空格又称为蒙古文空格。在蒙古文语料中蒙古文空格与普通空格之间的成分被标注为格的附加成分。蒙古文的词与词之间需要格的附加成分,格的附加成分本身仅有语法意义而无语义意义。蒙古文加上格的附加成分之后,蒙古文句子才能变得通顺[1]。蒙古文名词在句子中的语法成分由其后跟随的格的附加成分来决定。由于主格就是名词本身,无格的附加成分,因此格的附加成分共有 7 类,图 4 给出了蒙古文中的 7 类格的附加成分,共计 13 个。蒙古文名词与不同的格的附加成分之间由蒙古文空格连接充当不同的句子成分。但是在统计机器翻译的过程中如果不对格的附加成分进行预处理,Moses 会将蒙古文空格识别为普通空格进行处理,从而容易使一个蒙古文词从中间被切分开来而被识别成两个词甚至多个词。此情况一方面会造成数据稀疏并严重影响蒙古文-汉文统计机器翻译过程中的词对齐效果,另一方面会使蒙古文句子长度明显增长,影响翻译质量和最终的 BLEU 测评。因此,我们认为格的附加成分预处理(简称格处理)是十分必要的。对蒙古文语

料进行格处理是本文的研究重点。

在包含 571075 个蒙古文词干的语料中，4 种控制符共计出现 50518 次，其中窄宽度无间断空格出现 49921 次，出现频率非常高，说明格的附加成分对实验的影响较大。本文处理格的附加成分的方法主要有两种：1)将控制符去除，然后将格的附加成分与前面的词干进行连接形成一个新的词；2)将控制符与格的附加成分一同去除，只留下词干部分。两种格处理方法得到的语料词频分析如表 2 所列，在总词数方面，由于方法 1)将格的附加成分加到词干之后势必使原本相同的词干变成了不同的词，因此总词数增多；方法 2)将格的附加成分直接去除，使得词干提取更加精炼，从而出现更多相同的词干，因此总词数减少。在一次词频方面，方法 1)处理的词干语料中仅出现一次的词的频率比未经过格处理的蒙古文词干语料中仅出现一次的词的频率低 2.62%，方法 2)处理的词干语料中仅出现一次的词的频率比未经过格处理的蒙古文词干语料中仅出现一次的词的频率低 8.8%。数据证明两种方法都缓解了数据稀疏问题，但第二种方法更高效。因此本文采用将控制符与格的附加成分一起去除的格处理方法。

元音间隔符(如：ᠨ中间的空隙)	18	OE	(Unicode)	E1	A0	8E (UTF-8)
零宽禁连接符	20	OC	(Unicode)	E2	80	8C (UTF-8)
零宽度连接符(如：ᠬ想以中间形式开头)	20	OD	(Unicode)	E2	80	8D (UTF-8)
窄宽度无间断空格(如：ᠠᠷ前面的空隙)	20	2F	(Unicode)	E2	80	AF (UTF-8)

图 3　4 种控制符

1	2	3	4	5	6	7
定格	向位格	宾格	凭借格	从比格	和同格	联合格

图 4　格的附加成分

表 2　经过格处理的语料的词频分布

	蒙古文词干	第一种格处理方法	第二种格处理方法
总词数	34335	37065	27517
1	20818	23717	15906
2	4812	4882	3917
3	1988	1984	1662
4+	6715	6481	6031
一次词频占比/%	66.6	63.98 (−2.62)	57.8 (−8.8)

本文对蒙古文原始语料进行了第二种方法的格处理，数据显示经过格处理的语料总词数有所减少。去除格的附加成分的词中仅出现一次的词的频率比蒙古文词中仅出现一次的词的频率低 2%。虽然蒙古文中的格的附加成分有语法意义，对其进行去除会导致一定程度的信息丢失，但是数据证明经过格处理之后的蒙古文语料的数据稀疏问题得到了显著缓解；另一方面还可以提高语料的词对齐的准确率，因此假设所提出的对语料的格处理方法可以提高翻译质量。

1.4　蒙古文拉丁转写

蒙古文本身存在形同音异的现象，在人工搜集整理语料时会产生难以避免的拼写错误。为了弱化该错误，我们选择中间字符转换将蒙古文转化为拉丁字符。另一方面，拉丁转写还可以使得其在计算机处理的过程中能够采用编码方式替代原来的蒙古文字在通用的系统下运行，许多内部运行工作、存储工作都可以用拉丁转写形式进行，在需要时再转换为传统文字[8]即可。图 5 给出了经过拉丁转写之后的蒙古文词。

TVsO　TVsO+1A　TVsO+TAI

图 5　蒙古文词切分及拉丁转写

2　汉文语料预处理

中文分词技术是中文信息处理的一个重要的前置过程[9]，同时也是蒙古文-汉文统计机器翻译目标语言语料处理领域的关键技术和难点。中文分词可以有效地提高词对齐的效果，使得后续的统计机器翻译工作得以顺利进行。中文分词的理论研究可归纳为：3 种主要分词算法及组合算法研究、中文分词歧义消除、未登录词识别与分词和词性标注评测研究。目前已有较多分词算法，大致可以归纳为：词典分词方法、理解分词方法、统计分词方法、组合分词算法。

目前主要采用 Zhang 等人[10]提出的 ICTCLAS 中文分词系统对中文语料进行词切分。但是按词切分存在以下问题：黄昌宁等人[11]提出汉文语料按词切分的方法对未登录词方面的识别性能较差；陈晓等人[12]提出汉文语料按词切分的方法容易产生歧义从而对机器翻译产生干扰；奉国和等人[13]提出的汉文语料按词切分也会造成分词错误，从而在词对齐阶段引起对齐错误；Wu 等人[14]提出由于蒙古文-汉文机器翻译语料规模较小，通过词频分析发现按词切分会产生严重的数据稀疏问题。相比于传统的粗粒度的词切分方法，近年来学者们提出的细粒度的词切分方法在自然语言处理的多个应用领域得到了较好的效果[15-17]。而鉴于以上词切分产生的问题，我们认为在蒙古文-汉文统计机器翻译语料规模较小的情况下，中文语料预处理应采用细粒度的按字切分方法。本文基于中文语料进行了词频分析，中文语料库中词和字的频率分布如表 3 所列。

表 3　中文语料库中词和字的频率分布/%

词频	中文词	中文字
1	42.61	25.99
2	15.29	7.74
3	8.07	4.80
4	5.38	3.75
5+	28.65	57.72

结果显示，在频率小于或等于 4 次的情况按字切分语料比按词切分语料的百分比更低；在频率大于或等于 5 的情况，在按字切分的语料中占比为 57.72%，而在按词切分的语料中只有 28.65%。以上统计数据证明，粗粒度的词对齐相比细粒度的字对齐存在更严重的数据稀疏问题。

3　实验总结与分析

3.1　实验环境及语料介绍

蒙古文词切分处理阶段采用明玉等提出的基于词典、规则和统计相结合的方法，并在切分的基础上进行词干的提取

以及格附加成分的预处理。蒙古文拉丁转写阶段采用本文所提拉丁转写方法将蒙古文语料转写成拉丁形式。中文语料处理阶段采用 Zhang 等人提出的 ICTCLAS 中文分词系统对中文语料进行按字或词的切分。使用 Moses 作为统计机器翻译的系统，Moses 的配置如下：通过 Och 等[18]提出的 GIZA++进行词对齐；采用 Koehn 等[19]提出的基于短语的统计机器翻译解码方式进行解码；采用 Och 等[20]提出的 MERT 进行权值的优化，使用三元的语言模型。

统计机器翻译自动评价标准是进行统计机器翻译模型判别训练的必要条件，也是衡量机器翻译质量的重要指标[21]。本文实验的统计机器翻译评价标准采用 Papineni K 等[22]提出的 BLEU 评价标准。

本次实验选取了本领域目前唯一公开的经过长度筛选的 CWMT09 蒙古文-汉文统计机器翻译双语语料中长度在 50 以内的 65752 句双语日常生活语料作为训练集；为了调节合适的训练参数，从训练语料库中选择 1000 句语料建立了蒙古文-汉文统计机器翻译的开发集。由于本语料的测试集是未公开的，因此为了测试最终的翻译结果，我们建立了 1000 句测试语料。测试语料与训练语料不可以重复但是属于同一领域。实验语料信息如表 4 所列。

表 4 实验语料信息

	训练集	开发集	测试集
双语句对	64752	1000	1000
汉文语料规模	2.72MB	40kB	40.3kB
蒙古文语料规模	7.32MB	107kB	109kB
蒙文词总数	571075	7371	7555
汉文词总数	591521	8670	8792
汉文字总数	722099	10484	10556

3.2 实验结果的对比与分析

3.2.1 中文分词的实验结果

中文分词对比实验选择未经过处理的蒙古文原始语料和按词切分的汉文语料作为基线系统 1，选择蒙古文原始语料和按字切分的汉文语料作为基线系统 2。通过 Moses 进行统计机器翻译，实验结果如表 5 所列。基线系统 1 的翻译结果比基线系统 2 高 1.18 个 BLEU 点，证明了按字切分方法有效缓解了由按词切分方法带来的上述的问题，并且能够使翻译结果得到提升。

表 5 中文分词结果

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文原始语料	词	29.48
蒙古文原始语料	字	30.66(+1.18)

3.2.2 词干提取的实验结果

实验选择基线系统 1 和基线系统 2 作为基准实验系统，将经过词缀切分、词干提取之后的蒙古文语料与分别按词和字切分的汉文语料作为对比实验系统。通过 Moses 进行统计机器翻译，最终得出的实验结果如表 6 所列，实验结果显示对比实验通过统计机器翻译系统得到的翻译结果比基线系统 1 和基线系统 2 的翻译结果分别低 0.4 和 1.17 个 BLEU 点。正如 1.2 节提到的词缀的去除会导致蒙文信息的丢失，特别是在蒙古文-汉文统计机器翻译语料规模较小的情况下，由于词缀的切分、词干的提取造成了更严重的数据稀疏问题，验证了我们之前提出的假设。

表 6 词切分测试结果

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文词干	词	29.08(−0.4)
蒙古文词干	字	29.49(−1.17)

为验证本文提出的问题，对蒙古文语料进行词缀的保留，并统计了词频。词频分布情况如表 7 所列，表 7 中的数据显示保留一个和两个词缀的语料库中仅出现一次的词的频率比将词缀全部去除的语料中仅出现一次的词的频率分别降低了 9.44%和 10.22%。数据证明经过词干提取之后的蒙古文语料造成了较严重的数据稀疏问题。将词干提取的蒙古文语料与按字切分的中文语料作为基准实验系统，将保留一个词缀和保留两个词缀的蒙古文语料与按字切分的中文语料作为对比实验系统。通过 Moses 进行统计机器翻译最终得出的实验结果如表 8 所列，保留一个词缀的语料的翻译结果较基准实验系统的翻译结果提高了 0.22 个 BLEU 点，保留两个词缀的语料的翻译结果较基准实验系统的翻译结果提高了 0.57 个 BLEU 点。实验结果表明，进行词缀保留可以有效缓解由词缀去除带来的信息丢失和它造成的数据稀疏等问题，验证了我们的结论。

表 7 保留词缀语料词时的词频分布情况

	词干	一个词缀	两个词缀
总词数	34335	37248	38170
1	20818	21291	21521
2	4812	5264	5475
3	1988	2326	2459
4+	6715	8366	8714
一次词频占比	66.6%	57.16%(−9.44%)	56.38%(−10.22%)

表 8 保留词缀的实验结果

蒙古文语料处理方法	中文词切分方法	BLEU
词干	字	29.27
词干+第一词缀	字	29.49(+0.22)
词干+第一词缀+第二词缀	字	29.84(+0.57)

3.2.3 格处理的实验结果

为了验证语料格处理对实验结果的影响，分别对蒙古文原始语料和蒙古文词干语料进行格处理。通过 Moses 进行统计机器翻译，最终得出的实验结果如表 9 所列。实验结果显示，在蒙古文原始语料基础上进行格处理之后的语料与分别按词和字切分的汉文语料所组成的对比实验系统的翻译结果比基线系统 1 和基线系统 2 的翻译结果分别提高了 0.78 和 0.38 个 BLEU 点。在蒙古文词干基础上进行格处理之后的语料与分别按词和字切分的汉文语料所组成的对比实验系统的翻译结果比未进行格处理的蒙古文词干语料与分别按词和字切分的汉文语料所组成的实验系统的翻译结果分别提高 0.53和 1.46 个 BLEU 点。

表 9 蒙古文语料格处理的实验结果

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文原始语料	词	29.48
蒙古文原始语料	字	30.66
蒙古文原始语料+格处理	词	30.26(+0.78)
蒙古文原始语料+格处理	字	31.04(+0.38)
蒙古文词干	词	29.08
蒙古文词干	字	29.27
蒙古文词干+格处理	词	29.61(+0.53)
蒙古文词干+格处理	字	30.73(+1.46)

两组不同语料的对比实验结果均证明了前面提出的假设：经过格处理之后的蒙古文语料能够有效地缓解数据稀疏问题，提高蒙古文-汉文统计机器翻译的质量。

3.2.4　蒙古文拉丁转写的实验结果

为了验证蒙古文拉丁转写对实验结果的影响，在对蒙古文原始语料和蒙古文词干语料进行格处理的基础上又进行了拉丁转写。通过 Moses 进行统计机器翻译，最终得出的实验结果如表 10 所列。实验结果显示，在蒙古文原始语料的基础上先进行格处理再进行拉丁转写处理的语料与分别按词和字切分的汉文语料所组成的对比实验系统的翻译结果比进行格处理但未进行拉丁转写的蒙古文语料与分别按词和字切分的汉文语料所组成的实验系统的翻译结果分别提高1.72和1.64个 BLEU 点，比基线系统 1 和基线系统 2 的翻译结果分别提高 2.5 和 2.04 个 BLEU 点。在蒙古文词干语料的基础上先进行格处理再进行拉丁转写处理的语料与分别按词和字切分的汉文语料所组成的对比实验系统的翻译结果比进行格处理但未进行拉丁转写的蒙古文词干语料与分别按词和字切分的汉文语料所组成的实验系统的翻译结果分别提高 0.37 和 0.36个 BLEU 点，比基线系统 1 和基线系统 2 的翻译结果分别高出 0.5 和 0.43 个 BLEU 点。

表 10　蒙古文语料拉丁转写测试结果

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文原始语料＋格处理	词	30.26
蒙古文原始语料＋格处理	字	31.04
蒙古文原始语料＋格处理＋拉丁转写	词	31.98(＋1.72)
蒙古文原始语料＋格处理＋拉丁转写	字	32.70(＋1.64)
蒙古文词干＋格处理	词	29.61
蒙古文词干＋格处理	字	30.73
蒙古文词干＋格处理＋拉丁转写	词	29.98(＋0.37)
蒙古文词干＋格处理＋拉丁转写	字	31.09(＋0.36)

两组实验数据充分证明了拉丁转写之后能够区分出许多型同音异的蒙古文词，并且矫正了蒙古文中的拼写错误，减小了混淆的概率，进一步提升了翻译的质量。

3.2.5　其他语料的实验结果

为了证明所提方法的通用性，我们在本实验室目前整理的句子长度在 50 词以内的 15 万句混合领域蒙古文-汉文双语语料上进行了本文所提出的语料预处理方法的实验。从 15 万句语料中随机抽取 1000 句作为开发集，为了与本文实验进行对比，仍然选用 3.1 节所描述的日常生活语料的 1000 句作为测试集。由于篇幅有限，表 11 只列出具有代表性的 3 组重要实验数据，其他组实验数据的总体趋势相同。

表 11　实验结果总结

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文原始语料	词	13.12
蒙古文原始语料＋格处理	字	21.59
蒙古文原始语料＋格处理＋拉丁转写	字	**25.83**

由于语料为多领域混合语料，开发集也是从多领域语料中抽取的，而测试集仅来源于日常生活语料，因此本组实验的测评结果较低，从而仅作为语料预处理方法的证明实验，但实验的结果仍然清楚地表明了格处理方法以及拉丁转写的方法具有明显的优势。

3.3　实验总结

本文对比实验的语料预处理方法大体上涵盖了蒙古文-汉文统计机器翻译语料预处理方面最主要也最前沿的技术。通过基准实验系统与统计机器翻译结果的 BLEU 值的比较可发现，在中文语料方面按字切分能够有效缓解由小语料按词切分带来的对齐错误和模糊以及未登录词等一系列问题，提高了词对齐方面的准确率从而得到更好的翻译结果。在蒙古文语料方面，经过词缀的切分和词干提取，可以使原本复杂的蒙古文与汉文进行词对齐时的难度降低，但由于蒙古文语料规模较小，本身存在较严重的数据稀疏问题，词干的提取中存在较多的格的附加成分，加重了数据稀疏问题从而影响了翻译质量；蒙古文语料格处理缓解了数据稀疏问题，使得翻译质量得到显著提升；经过拉丁转写的蒙古文语料能够更容易地区分出型同音异的蒙古文词，并且矫正了蒙古文中的拼写错误，减小了混淆的概率，大大提升了翻译的质量。

本文提出的在蒙古文原始语料的基础上直接进行格处理再进行拉丁转写的语料和中文按字切分的语料所组成的对比实验系统与基线系统 1 和基线系统 2 进行对比的实验结果如表 12 所列。本文提出的语料预处理方法得到的统计机器翻译的结果比基线系统 1 的翻译结果提高 3.22 个 BLEU 点，比基线系统 2 的翻译结果提高 2.04 个 BLEU 点，取得了令人满意的翻译效果。并且通过较大规模的语料实验也证明了所提语料预处理方法的优越性。

表 12　实验结果总结

蒙古文语料处理方法	中文词切分方法	BLEU
蒙古文原始语料	词	29.48
蒙古文原始语料	字	30.66
蒙古文原始语料＋格处理＋拉丁转写	字	**32.70**

结束语　本文总结了当下蒙古文-汉文统计机器翻译语料预处理的主要方法，并在此基础上对蒙古文语料提出了格处理的方法。通过组合不同方法处理的语料共得到 20 组对比实验系统。通过实验结果分析了各种语料预处理方法对蒙古文-汉文统计机器翻译的影响，证明了本文所提出的格处理方法使蒙古文-汉文统计机器翻译质量得到显著的提升。并且提出了一种新的语料预处理方法：中文语料按字切分，在蒙古文原始语料的基础上对蒙古文语料进行格处理再进行拉丁转写。本文提出的语料预处理方法的翻译结果比基线系统 1 高3.22个 BLEU 点。另外，为了证明本文提出的语料预处理方法的通用性，在较大规模的多领域混合语料上进行相同的语料预处理，并通过实验证明了本文所提方法的通用性。我们还对实验结果中的一系列问题进行了分析和总结，并通过实验对其进行证明。

后续我们将会继续研究蒙古文-汉文统计机器翻译语料预处理对翻译的影响，在所提方法的基础上继续进行优化。词干提取方面，我们需重点研究如何降低数据稀疏问题以及词缀的去除导致的信息丢失等问题，并在此基础上结合格处理方法，争取能够得到更好的蒙古文-汉文统计机器翻译语料预处理方法，提高翻译质量。

我们还会在今后的研究中继续收集整理蒙古文语料，以得到较大规模的高质量的蒙古文-汉文统计机器翻译语料，并且在大规模的语料方面继续采用本文所提出的语料预处理方法，进而证明所提方法的优越性。

另一方面，我们会在目前取得较好成果的神经网络机器翻译系统中应用提出的语料预处理方法，以得到较好的蒙古文-汉文机器翻译结果。

参 考 文 献

[1] NICOLAI G, KONDRAK G. Leveraging inflection tables for stemming and lemmatization[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016:1138-1147.

[2] NA S W. Mongolian word root, stem, suffix automatic segmentation system[J]. Journal of Inner Mongolia University (Humanities and Social Sciences Edition), 1997(2):53-57. (in Chinese)
那顺乌日图. 蒙古文词根、词干、词尾的自动切分系统[J]. 内蒙古大学学报(人文社会科学版), 1997(2):53-57.

[3] SINGH J, GUPTA V. Text Stemming: Approaches, Applications, and Challenges[J]. ACM Computing Surveys (CSUR), 2016, 49(3):45.

[4] WU J, HOU H X, BAO F L, et al. Template-based model for BiRNN Mongolian-Chinese machine translation[C]// Proceedings of TAAI 2015. 2015.

[5] HOU H X, LIU Q, NA S W, et al. Mongolian Word Segmentation Based on Statistical Language Model[J]. Pattern Recognition and Artificial Intelligence, 2009, 22(1):108-112. (in Chinese)
侯宏旭，刘群，那顺乌日图，等. 基于统计语言模型的蒙古文词切分[J]. 模式识别与人工智能, 2009, 22(1):108-112.

[6] ZHAO W, HOU H X, CONG W, et al. Research on Conditional Random Fields Based Mongolian Word Segmentation[J]. Journal of Chinese Information Processing, 2010, 24(5):31-35. (in Chinese)
赵伟，侯宏旭，从伟，等. 基于条件随机场的蒙古文词切分研究[J]. 中文信息学报, 2010, 24(5):31-35.

[7] MING Y. Researching of Mongolian Word Segmentation System Based On Dictionary, Rules and Language Model[D]. Hohhot: Inner Mongolia University, 2011. (in Chinese)
明玉. 基于词典、规则与统计的蒙古文词切分系统的研究[D]. 呼和浩特：内蒙古大学, 2011.

[8] 申晓亭. 少数民族文字拉丁转写的意义与方案[C]// 全国少数民族语言文字信息处理学术研讨会. 2007.

[9] XU J J, SUN X. Dependency-based gated recursive neural network for Chinese word segmentation[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016:567-572.

[10] ZHANG R, YASUDA K, SUMITA E. Improved statistical machine translation by multiple Chinese word segmentation[C]// Proceedings of the Third Workshop on Statistical Machine Translation. Ohio: Association for Computational Linguistics, 2008:216-223.

[11] HUANG C N, ZHAO H. Chinese Word Segmentation: A Decade Review[J]. Journal of Chinese Information Processing, 2007(3):8-19. (in Chinese)
黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报, 2007(3):8-19.

[12] 陈晓，靳光瑾，黄昌宁. 基于字的分词方法的实验研究：第九届全国计算语言学学术会议[C]// 全国计算语言学学术会议. 2007:52-57.

[13] FENG G H. Review of Performance Evaluation of Text Classification [J]. Journal of Intelligence, 2011, 30(8):66-70. (in Chinese)
奉国和. 文本分类性能评价研究[J]. 情报杂志, 2011, 30(8):66-70.

[14] WU J, HOU H X, LI J T, et al. Adapting Attention-Based Neural Network to Low-Resource Mongolian-Chinese Machine Translation[C]// International Conference on Computer Processing of Oriental Languages. Kunming, China: Springer International Publishing, 2016:470-480.

[15] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin (Germany): Association for Computational Linguistics, 2016:1715-1725.

[16] LEE J, CHO K, HOFMANN T. Fully Character-Level Neural Machine Translation without Explicit Segmentation[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016:1693-1703.

[17] PRABHU A, JOSHI A, SHRIVASTAVA M, et al. Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text[J]. ArXiv Preprint ArXiv: 1611.00472, 2016.

[18] OCH F J, NEY H. A systematic comparison of various statistical alignment models [J]. Computational Linguistics, 2003, 29(1):19-51.

[19] KOEHN P, HOANG H, BIRCH A, et al. Moses: Open source toolkit for statistical machine translation[C]// Proceedings of the Association for Computational Linguistics. Prague (Czech Republic): Association for Computational Linguistics, 2007.

[20] OCH F J. Minimum error rate training in statistical machine translation[C]// Proceedings of the Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003:440-447.

[21] YANG N. Neural Network Learning for Statistical Machine Translation[D]. Hefei: University of Science and Technology of China, 2014. (in Chinese)
杨南. 基于神经网络学习的统计机器翻译研究[D]. 合肥：中国科学技术大学, 2014.

[22] KOEHN P, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002:311-318.