基于统计机器翻译技术的胡都木—托忒文本转写的实现*

伊·达瓦　王羡慧　米尔阿迪力江·麦麦提**

（新疆大学　乌鲁木齐　830046）

[内容提要] 针对蒙文多文种文本（如传统蒙文 TM，新蒙文 NM 及托忒文 Todo）的互换显示需求，本文研究了基于短语的统计机器翻译技术的自动转写方法。首先，人工建立上述三文平行 6 万条句对语料。其次，利用 TM 和 Todo（NM）双文句对中，词间空格信息对，TM 功能词与前词强制连接，生成双文句及词对齐语料，并生成统计翻译模型和语言模型。最后，借助于 Moses 解码器实现双文的自动转写。实验分别用 300 开发句和测试句进行 TM－Todo 句文双向互译时，其 BLEU 值分别达到了 57.82% 和 58.03%，比先前汉－蒙语机器翻译最好 BLEU 值：29.86%，近高一倍。

[关键词] 蒙文胡都木托忒文本　平行语料　统计机器翻译　Moses　余弦相似度

[中图分类号] H2　[文献标识码] A　[文章编号] 1674－3067(2014) 02－0062－10

1. 引言

蒙古语属于多文种，多方言复杂语言范畴。如图 1 所示，它是一种因地区不同，国家不同而使用不同文字（字符）结构的书面语言。它不仅各文字构词以及成句规律有区别，而且各文字的发音规律也不一致。目前有的地区甚至同时并用两种形式的书面语言电子媒体[1]。与汉语，英语及阿拉伯语文字系统信息处理环境相比较，由于蒙文的垂直排版编译特点，使用字符种类多及技术资源短缺，通用系统软件尚未完全支持蒙文信息处理，蒙文多文种 Unicode 代码标准尚未完善等原因，蒙文的信息化进程相对而言较滞后。另外，现用蒙文录入显示市场软件种类较多，字库代码设计标准不统一，一般相互不兼容。这造成了蒙文电子化办公效率低，信息通信不疏通等较为突出的问题。特别是，居住在新疆地区的蒙古族群众，由于教育领域用 H（Hudum）或 TM（TraditionalMongolian 内蒙地区现用文字系统）文字媒体，而社会机构（出版界，政府办公）使用 H 和 T（Todo）两种文字信息媒体，蒙文网络通信用 TM 文字或 NM（New Mongolian 蒙古国现用斯拉夫文字）文字传递信息。对于本来离现代文化较偏远的民众，这使得他们因语言环境障碍不得不放弃古老文化而随波逐流。

* 新疆维吾尔自治区自然科学基金（2012211A012）。

** [作者简介] 伊·达瓦，博士，教授（博导），研究方向：计算语言学，人工智能学，自然语言信息处理及机器翻译。王羡慧（1980—），通讯作者，博士，副教授。米尔阿迪力江·麦麦提，研究生。

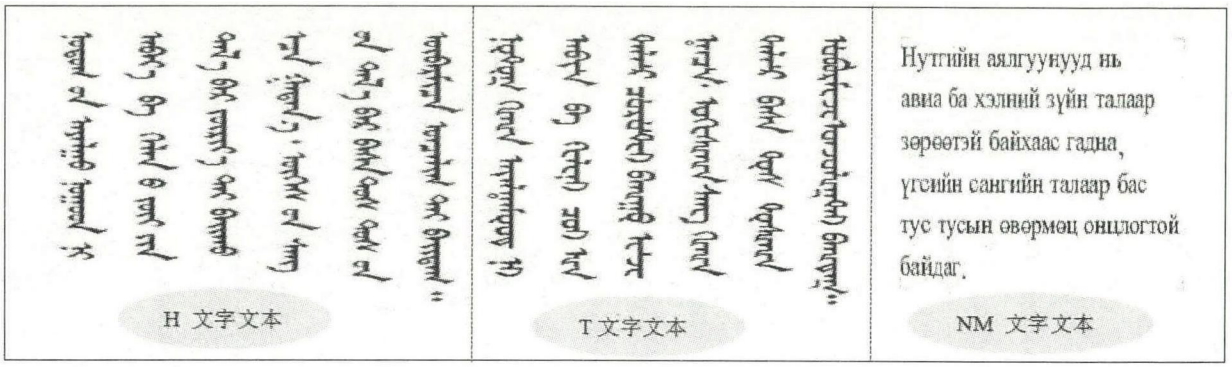


图 1　蒙文常用电子化文本样本

为了改善新疆地区蒙古族群众使用语言文字信息化工作的现状，并促进其进一步发展，新疆大学自治区多语种技术重点实验室，在 2011 年及 2012 年，先后争取到了新疆自治区科学技术厅自然科学基金项目和新疆自治区科技援疆项目。目前主要在以下两个方面开展基础的研究工作：

(1)基于开源 OpenType 技术的蒙古文办公软件的实现。

(2)研发蒙文多文种媒体信息的自动转写通信方法和技术。先行相关研究的初步成果，已在国内外相关学术会议及期刊上发表[2,3,4]。为了本技术的后续产品化开发，本文分析了实现相关技术的基础理论依据以及近期的研究成果。

本文在第 2 节简述蒙文信息处理存在的相关问题。第 3 节概括介绍了机器翻译技术，第 4 和第 5 节介绍该系统的实现及实验测试，在最后的第 6 节给出结论以及今后工作的重点。

2. 蒙文及信息化通信若干问题

2.1. 蒙文信息处理现状及存在问题

蒙文属于阿勒泰语系蒙古语族。蒙古语族及突厥语族 TBL(土耳其语，维哈柯语)语言均属于信息处理复杂语种范畴。相比于 TBL 语族语言，蒙文的信息化处理的难度更大。主要原因有：①现用蒙古文国际标准编码字符集中只收录了蒙文的“名义字符”标准，而没有收录显现字形标准。新疆地区使用托忒文字和锡伯文字名义字符标准也有，但不全。微软的 Vista 操作系统近期才支持蒙古文的“显现字形”变换处理，新版 Office2007 现支持蒙文的竖排排版。目前 Vista 中也提供了蒙文输入法，但是这类输入法只针对传统蒙文设置[5,6]；②现尚未制定 Unicode 标准托忒文输入法；③托忒文键盘输入拉丁字母标准尚未制定；④蒙文多文种文本(text)转换处理技术途径尚未认真得到研究；⑤现用蒙文处理软件种类多，由于标准不统一，相互不兼容；⑥现用蒙文文字信息处理软件智能化程度低，技术落后，过于依赖语言学的书面知识；⑦因蒙文信息处理技术力量薄弱，电子化资源的保护管理意识不够高，语言文字数字化资源严重缺少等多种原因，目前全国范围内蒙文信息处理通信事业的发展比较滞后。尤其是新疆地区使用蒙文的信息化开发研究工作几乎处于空白状态。

2.2 蒙文语言文字信息化及网络通信相关问题

图 1 所示为现用蒙文文字系列中部分的代表性样本。实际上蒙文信息处理涉及到如图 2 所示的多种文字字符数据。所以，面向蒙文的信息处理，网络通信的研究实际上是个国际性研究课题。需要各方齐心协力，资源共享，技术合作。

图 2 蒙文电子化数据实用样本

比方说 Todo 文中一词[ᠵᠢᠷᠭᠠᠯ](见图 3 右上角),其读音是[jirghal],而录入形式为/jirgal/,而 unicode 代码形式为(jirg'al),与该词同意的其他蒙文词,在系统中所表现出来的形式,显然是有差异的,互换技术的实现上不可能用同一字符串来进行替换或者通信。假如计算机用同一个字符串的代码录入,或者用同一个代码形式传递该词,在对方接受端都会出错误或者不能正确显示。

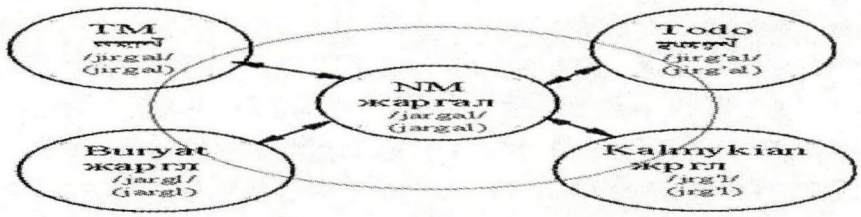


图 3 Todo 文一词[ᠵᠢᠷᠭᠠᠯ]在其他蒙文词中表现形式

另外,比如在键盘上录入的一个 TM 文句子/agvla yindegere/,如果直接用录入代码串形式转换为 Todo 文句子,不可能获得 Todo 文句子/uula in deere/的代码串,更不可能得到 Todo 文短语 ᠬᠤᠤᠯᠠ ᠤᠨ ᠳᠡᠭᠡᠭᠡ。甚至代码串/agvla yin degere/,用计算机语音合成技术实施发音,也绝不可能发出/uula in deere/的语音流。

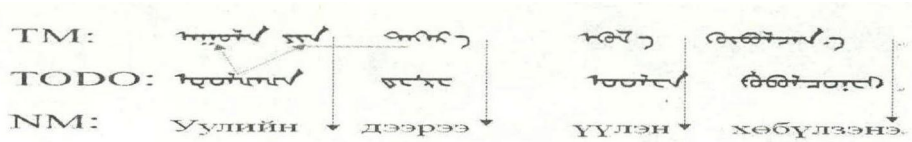


图4 常用蒙古多文种句对关系举例

总的来说,对于如图4所示那样的现用蒙文文本之间的转换处理及通信,既不能用字符单位一一对应互换,又不能用一词或空格区分字符串单位对应代替。这需要智能化的信息处理技术,才能实现各文本的正确互换处理。为此,该文提出了基于短语的统计机器翻译技术的蒙文多文种的转写处理方法。

2.3 相关研究现状

目前,有关蒙文多文种媒体信息转换处理方面的研究比较少。在文献[7]中,日本筑波大学 Ishikawa 研究组研究了基于语法规则的 TM 文和 NM 文双向转写方法。该文只涉及 TM 和 NM 文处理,没有研究 Todo 文及其他文字文本的转写处理问题。该文侧重语法规则法,尝试了 TM 和 NM 文文本转换实验。蒙古国 Y. Namsurai's 研究组,开展了基于语法规则及用少量 NM 文特定小说数据尝试 NM 文到 TM 文的转写实验[8]。近年来,基于统计机器翻译的汉 - 蒙(TM)机器翻译[9,10]或日 - 蒙(TM)机器翻译的研究[11,12]也不少。而对于蒙 - 蒙文间的机器转写或者机器翻译的相关研究则极少。本研究在先行的文献[2,3,4]中,也尝试了基于数据库及基于语言学规则方法的蒙文多文种的转换处理实验。

3. 统计机器翻译技术

随着计算机网络的快速普及应用,多语种文字网页信息已进入到了人民的生活中。为了能最大限度地利用好网页信息,多文种语言文本信息需要进行翻译。当然,通过人工方式翻译各文种网页信息是不现实的。为此,到了今天的大数据时代,基于计算机的自动翻译技术的实现是多语言全球化通信的又一个新的梦想。

目前基于计算机的机器翻译方法分为基于规则的翻译方法 RBMT(Rule based MT),基于实例的翻译方法 EBTM(Example Based MT)以及基于大型语料库和统计技术的机器翻译方法 SMT(Statistical MT)等三大技术模式。

· RBMT 翻译模式,只能按已知的语言学规则转换翻译有限的,语法规则较严密的文本,对于千变万化的自然语言一一描述其规则是不现实的。

EBMT 翻译方法也是一种基于语料库的方法。这种方法只对比实例句子,通过实例对比原理进行翻译[13]。对于活生生的自然语言句子成分跟踪设立实例库也有难度。

SMT 翻译方法把机器翻译看成是一个信息传输的过程,用一种信道模型对机器翻译进行解释。该方法认为:源语言句子到目标语言句子的翻译是一个概率问题,任何一个目标语言句子都有可能是任何一个源语言句子的译文,只是概率不同而已。具体方法是将翻译看做对原文通过模型转换为译文的解码过程。互译时要找到概率最大的句子。因此,SMT 法不需要任何语言学知识,实例模板等。在平行语料规模足够大的条件下,一般可以获取较好的译文质量。对于语序语法规则较接近的语言,由于 SMT 技术可以避免复杂的调序处理,互译效果明显提升[14,15]。

针对蒙文各文本语序相同,语法结构一致的特点,本文使用基于短语的统计机器翻译方法,引用通用 Moses 解码器[16]实现蒙文多文种文本的转写。

4. 翻译系统的构造及工作原理

4.1 系统的构造及工作原理

基于统计翻译的实验系统主要由训练语料(包括双语平行语料库(Bilingual Corpus)和目标语言语料库(Target Language Corpus)两部分)、基于对齐语料的翻译模型(如,TM Model (Translation Model))、目标语言统计语言模型 LM(Language Model)及解码器(Decoder)等五个模块组成(见图.5)。翻译的实现过程简述如下:首先,研究开发蒙文多语种句子平行文本语料,即 Bilingual Corpus。本研究得到蒙文媒体出版发行单位的协助,邀请蒙文专家,人工建立了 6 万个句子的 NM – TM – Todo 平行语料(样品数据见图 6 所示),并参考 TM 文功能词(见表_1)信息,据双语句子长度信息对互译语言实施词对齐处理,生成了句对语料。

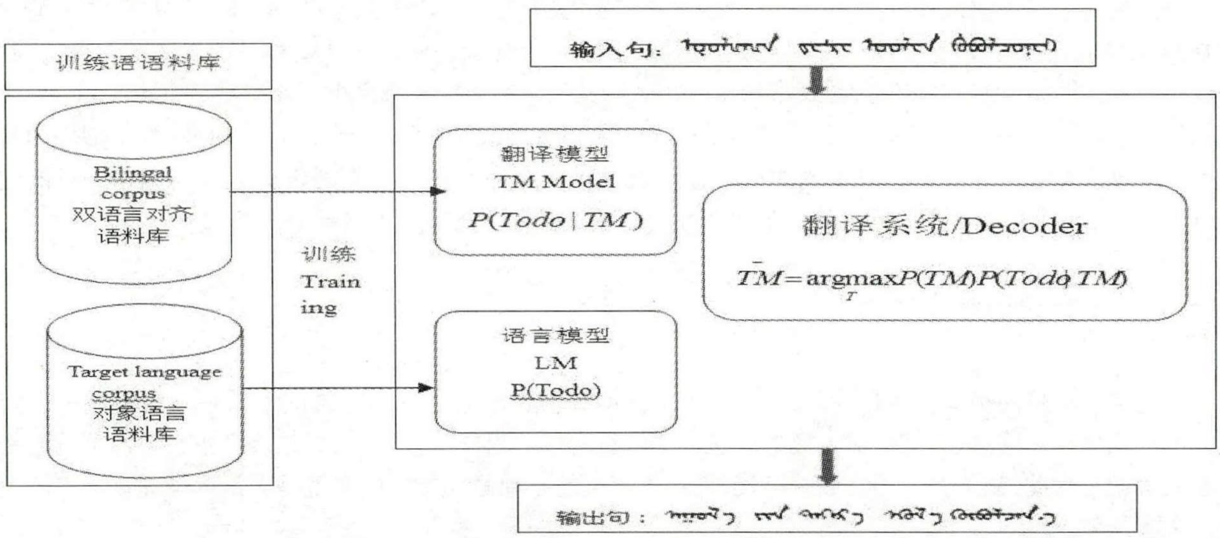
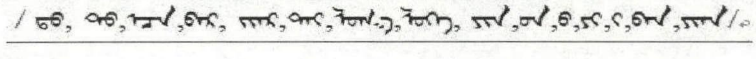


图 5 基于短语的统计翻译的蒙文多文种转写系统

图 6 蒙文多文种句对语料样本

表 1 TM 文功能词列表

a function words set of TM text system



其次,建立目标语言大规模文本语料。然后,利用双语对齐语料,借助于 Moses 软件训练从原语言 f 到目标语言 e 的统计翻译模型。同时,用目标语言文本数据(比如 TM 文数据),借助于开源软件 SRILM 训练 N – gram 统计语言模型,如 P(TM)。最后,利用 Moses 解码器实现对原语言 Todo 文句子(如 )实施翻译。解码时,首先对于输入句子选取若干个匹配较好的互译句

子 ek(k = 1..K),然后在目标语译选句子中选拔匹配几率最高的短语片并进行排序组合后作为最终译文输出(如,TM 文句子 [illegible])。对于原语言 Todo 到目标语言 TM 的统计翻译数理模型可以下式(1)表示,即

$$\hat{Todo} = \arg\max_{Todo} P(TM \mid Todo)P(TM) \qquad (1)$$

4.2 句子短语片的排列(调序)

如图 7 所示,在基于短语的机器翻译系统中,用空格区分的一个以上词组合的词串作为最小翻译单位(短语片)实施语言 f 和语言 e 的互译。虽然平行语料双语句子是对齐的,但是句子中短语片之间往往不是按语言 f 和语言 e 中词出现的顺序对齐的。因此,系统对源语言每个短语片进行互译完毕后,还需要按目标语言语法规律对已译短语片重新排序组合而生成目标语言句子。这个过程在统计机器翻译系统中叫做译文短语的调序。在统计机器翻译系统中,互译语言之间的语序的一致性直接影响系统最终的译文质量。所以,自动调序操作在统计机器翻译研究中是一个很重要的,而且很复杂的研究内容。基于短语的统计机器翻译系统中,考虑短语片出现顺序的翻译模型可以表示为式(2)。

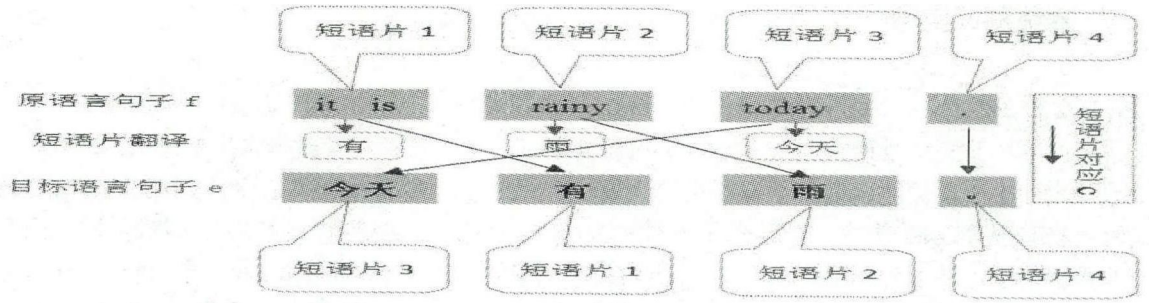


图 7 在短语对齐机器翻译中短语调序方法

$$P(f,c \mid e) = P(c \mid e)P(f \mid c,e) \approx P(c_1^I \mid e)\prod_{i=1}^{I} P(\overline{f_{c_i}} \mid \overline{e_i}) \qquad (2)$$

其中,$\bar{f},\bar{e}$ 分别是源语言和目标语言短语片,$C_1^I = c_1, c_2, \cdots c_i$ 表示短语片顺序,c_i 表示在目标语侧第 i 个短语片 $\bar{e}_i$ 所对应源语言侧的短语片的序号。$P(c_1^I \mid e)$ 称为调序模型。比如在图 7 所示英 - 中文句对中,原语言(英)各短语片的翻译短语顺序为/有雨今天。/,经调序处理后输出目标语言句子为/今天有雨。/。即,源语言句子中短语片 3 所对应的翻译短语片调序到了目标语言句子的第一个短语片位置。其他情况用类似方法进行调序。常用调序模型算法为 LBO(Lexicalized Block Orientation)模型,表示为下式(3),即,

$$class(c_i, c_{i+1}) = \begin{cases} monotone & (c_{i+1} = c_i + 1), \\ swap & (c_{i+1} = c_i - 1), \\ discontinuous & (\text{others}). \end{cases} \qquad (3)$$

如果在目标语侧两个短语片位置相邻,而且顺序与原语言短语片相一致时选择 monotone 类;在目标语侧两个短语片位置相邻,而且顺序与原语言短语片位置相反时选 swap;而两个短语片位置不相邻,而且为偏离时选 discontinuous 类进行调序。LBO 模型用上述 3 类近似式(2)为各短语片调序的概率之和。即,

$$P(c_1^I \mid e) \approx \prod_{i=1}^{I} P(class\ (c_i, c_{i+1}) \mid \bar{e}_i, \bar{e}_{i+1}) \qquad (4)$$

4.3 蒙 - 蒙语序调序

蒙文多文种句子虽然语序基本一致,但是,互译语言词与词间不是用空格做对齐的。主要是 TM 文句子中功能词不能与前词项像 Todo 和 NM 词那样进行语法连接。因此,本文首先,在 TM 文本语料中提取出所有功能词(见表 1);其次,再用互译句对中空格数判断句对长度是否相等。如果句对长度相等,那么互译句子中词与词间用空格区分对齐处理,否则,对 TM 句子(如图 8 那样),强制功能词与前词项用特殊符号(比如//)做无空格连接,使得源语言句子语序尽可能地调整为与目标语言一致。这样调序模型公式(4)按 monotone 类进行调序。即,句对按空格顺序 $c_{i+1} = c_i + 1$ 排序。

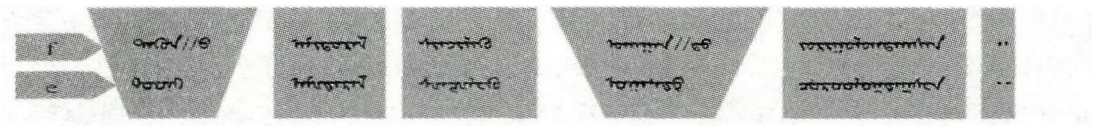


图 8 f 和 e 双语句子用功能词强迫空格对齐处理

5. 系统测试实验

5.1 实验条件

系统评测实验利用本研究开发的蒙文 NM,TM 及 Todo 三文字人工选择录人的 6 万条平行语料。对于短语表的生成及调序模型的学习用相同语料。考虑到蒙文多文种是同语序,同语法的相似性语言及试验语料有限等因素,本次实验设置了三组实验。(1)利用 5 万词的蒙文多文种 - 汉标注电子词典,对于源语言句子空格区分词项(字符串),检索目标语言匹配词项生成目标语言句子,再引用 BLEU 值和余玄相似法考察互译句对的相似性;(2) 以标准的基于短语的统计机器翻译系统作为参考,仅用双语句对数据,作为基线系统;(3) 对于 TM 句子进行功能词与前词项强制连接。实验语料细节见表 2。

表 2 本研究用语料

平行语料	句对文本	词条数	词数
学习集(TM)	6 万	520,000	35,430
学习集(Todo)	6 万	494,500	33,540
开发集(TM)	300 句	3,300	1,800
开发集(Todo)	300 句	3,542	1,500
测试集(TM)	250 句	2,600	1,400
测试集(Todo)	250 句	2,422	1,240

5.2 实验 -1:基于词典转写

5.2.1 实验方法:利用如表_3 所示 5 万词条的蒙 - 中文词标注电子词典,做两次 TM - - > Todo 方向句子转写测试:①对于源语言 TM 句子中的每个用空格区分词项,通过词典检索匹配 Todo 词,再与 TM 句子同序排列输出 Todo 句子。②对于源语言 TM 句子中的每个用空格区分词项,通过词典,引用下式(7)余弦相似算法匹配计算相似度量高的 Todo 词,再与 TM 句子同序排列输出 Todo 句子。

表3　蒙－汉对齐电子词典样本(单语5万词条)

30728|||чамар|||/chamar/|||[<Ne>]|||[illegible]|||/chamar/|||[illegible]|||/chimar/|||[名/ming2/：油茶面/you2 cha2 mian4/]
30729|||чамарлагдах|||/chamarlagdah/|||[<Ve>]|||[illegible]|||/chamarlagdahu/|||[illegible]|||/chimarlagdahu/|||[动：
30730|||чамарлалцах|||/chamarlaltsah/|||[<Ve>]|||[illegible]|||/chamarlaltsahu/|||[illegible]|||/chimarlaltsahu/|||[动：
30731|||чамарлах|||/chamarlah/|||[<Vt>]|||[illegible]|||/chamarlahu/|||[illegible]|||/chimarlahu/|||[动/dong4/：做油茶面
30732|||чамарлацгаах|||/chamarlats$a:h/|||[<Ve>]|||[illegible]|||/chamarlats$a:hu/|||[illegible]|||/chimarlatsa$a:hu
30733|||чамарлуулах|||/chamarlu:lah/|||[<Vt>]|||[illegible]|||/chamarlu:lahu/|||[illegible]|||/chimarlu:lhu/|||[动：]面
30734|||чамархай|||/chamarhai/|||[<Ne>]|||[illegible]|||/chamarhai/|||[illegible]|||/tsamarhi:/|||[名/ming2/：鬓角/bin4 jiao
30735|||чамархах|||/chamarhah/|||[<Vt>]|||[illegible]|||/chamarhahu/|||[illegible]|||/chimarhahu/|||[动/dong4/：嫌少/xia

5.2.2 Cosine 相似尺度

设有两个 n 和 m 维向量 A 和 B,如公式(6)所示,这两个向量的相似性由公式(7)给出。当 cosineθ =1,(θ =0^0)时,两个向量 A 和 B 相同,即 A 和 B 完全相似;当时,两个向量 A 和 B 完全不相同,即 A 和 B 无相关性;用 cosineθ 在[0,1]之间的取值度量两个向量 A 和 B 的相关程度[18]。

$$\begin{cases} A = a_1, a_2, \cdots a_n \\ B = b_1, b_2, \cdots b_m \end{cases} \quad (6) \qquad sim = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{m} (B_i)^2}} \quad (7)$$

5.2.3 BLEU 尺度:是互译文自动评估的常用尺度,依据 N－gram 匹配率由式(8)表示,

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log p_n\right) \quad (8)$$

其中,$p_n = \frac{\sum_i \text{译句 } i \text{和参照句 } i \text{中一致的 } N-gram \text{ 数}}{\sum_i \text{译句 } i \text{中全 } N-gram \text{ 数}}$, $BP = \begin{cases} 1 & if\ c \geq r \\ \exp(1-r/c) & if\ c < r \end{cases}$

C 为 MT 译句长度，r 为参考译句中最近译句长度。

5.2.4 实验－1 结果

上述实验－1 两种方法对源语言 TM 测试集300 句子转写输出 Todo 文300 句子，其结果与 Todo 文250 个测试句算出 BLUE 值和句子相似度 sim。其评估结果为表4 所示。

表4　实验－1 基于词典测试结果

方法	BLEU(%)	sim
①	35.6	0.62
②	37.2	0.68

5.3 实验－2:基于短语的统计翻译转写

本次实验测试 TM↔Todo 双向翻译。评测精度利用 BLEU 值和句子相似度 sim 分别进行评估。实验数据与表2 相同。基线测试用 TM 文功能词与前词项不做强制连接,而本文所提方法用 TM 功能词做与前词做强制连接后的句对语料分别进行。

为了增大目标语言句子的变异性,在获取 N－best 翻译结果时,只选取得分最高候补句子选入 N－best 中。目标语言 N－gram 语言模型的学习,除了使用表2 所示语料外,还增加了大量不同领域用词电子文本。引用开源软件 SRILM 工具,加入 Kneser－Ney 平衡参数进行1－5 元语言模型的训练。其中3－元模型训练设定为:// ngram－count －texttrain. tex －lm lm －interpolate －kndiscount1 －kndiscount2 －kndiacount3 －order 3.

解码器相关参数的设定如表5 所示。对于解码器各参数权重学习使用开发集,实施 Minimum Error

- Rate Training 方法学习。

表 5 评测参数设定

项目	条件
解码器	Moses（3）
翻译候补数 beam	10,20,50,100,200,500,1000
table - limit	20
调序	msd - bidirectional - fe
语言模型	1 - 5gram

表 6 基于短语机器翻译评测结果

BLEU(%) / 方向	baseline	proposed	sim
TM→Todo	53.26	57. 82	0.82
Todo→TM	54.87	58.03	0.87

表 6 中给出基于短语的统计机器翻译实验_2 的评测结果，图 9 为 Todo→TM 文转写实演。比较实验结果表 4 和表 6 可以看到，本文所提统计机器翻译方法的最好 BLEU 值（Todo→TM）58.03% 高于基线（baseline）值 54.87% 约 3.16%。比基于词典的测试 BLEU 值高出 20.83%，而译句与人工译句的相似度 0.87 也高于基于词典结果 0.68 近 0.2。译句与目标语测试集的相似性大幅度提高。另外，本文所提方法评测结果明显高于先行汉 - 民机器翻译研究目前最好的翻译结果 29.86%，高了近两倍[9]。这完全反映了基于短语的统计机器翻译对语序一致性语言文本翻译的有效性。

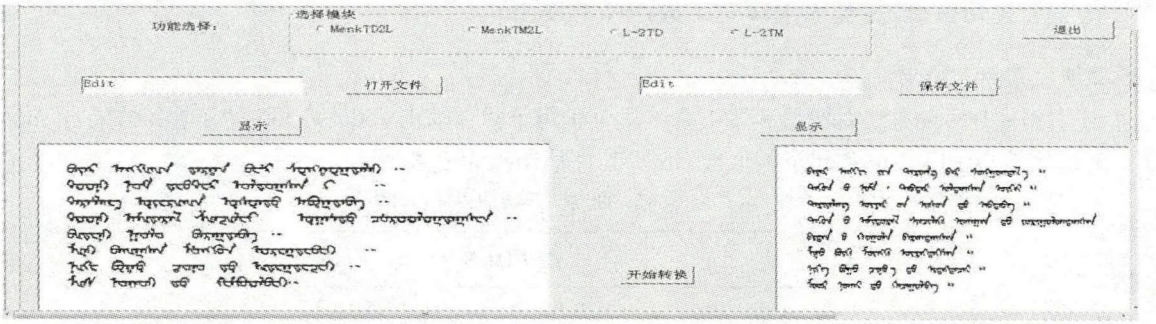


图 9 本研究开发系统 Todo - TM 文转写实演结果

6. 结 论

通过本文所提方法的实验结果分析发现，基于短语的统计机器翻译方法对于蒙文多文种文本的转换处理极为有效。尤其是，利用 TM 文功能词信息，针对 Todo（NM）文句子做词对齐处理后的短语统计翻译结果明显好于直接利用平行句子的互译效果。同时，实验结果也显示，基于词典的互换方式明显差于统计翻译效果。这完全反映了统计机器翻译对于语序差异性较小语言的互译优势。通过本次实验我们也会体会到基于短语统计机器翻译方法对于诸多相似语言（如维哈柯语言）间的转换处理将是一个有效地，便利快捷的方法。

由于现有蒙文多语种对齐语料有限，多文种平行语料的收集整理，人工翻译及预处理等先行研究尚未完善，因此，影响了本次实验结果。扩大多文种平行语料资源规模，改善现有语言资源质量，提升系统

性能是本研究今后工作的重点。

参考文献:

1. Idomucogiin Dawa, Satoshi Nakamura, A Study on Cross Transformation of MongolianFamily Language, Journal of Natural Language Processing,J - STAGE, Vol. 15 No. 5,2008, pp3 - 21.

2. I. Dawa, U. Aishan, A. Kahaerjiang,Y. Turgen, Creating and Analysis of a POS Tag Chinese - Mongolian Multi - Version Dictionary, International Journal of Computer Mathematics, 2014. 3 Vo. 3(16), 123 - 127.

3. 王玲,达瓦·伊德木草,吾守尔·斯拉木, 维哈柯及蒙文多文种语言相似性考察研究同语系多种黏着语言相似性研究,中文信息学报,2013,Vol. 27(6), 180 - 1186.

4. 达瓦·伊德木草,木哈亚提·尼亚孜别克,吾守尔·斯拉木,语音技术在少数民族语言的应用研究,新疆大学报,2014, V0l. 31(1),88 - 96.

5. 白双成,张劲松,呼斯勒,蒙古文输入法输入码方案研究,中文信息学报, 2013,27(6),169 - 173.

6. 国家质量监督检验检疫总局,国家标准化管理委员会 GB256914 - 2010. 信息技术传统蒙古文名义字符,变形显现字符和控制字符使用规则 S]. 北京. 中国标准出版社,2011. 11.

7. T. Ishikawa,. et al. A Bidirectional Translation Method for the Traditional and Modern Mongolian-Scripts//. Proceeding of the Eleventh Annual Meeting of The Association for Natural Language Processing. 2005: 360 - 363.

8. Y. Namsurai. et al. The database Structure for BI - Directional Textual Transformation Between Two Mongolian Scripts//, Proceeding ICEIC2006,265 - 268.

9. 王斯日古楞,斯琴图,那顺乌日图, 汉蒙统计机器翻译中调序方法研究, 中文信息学报, 2011, 25(4): 88 - 92.

10. 银花,王斯日古楞,艳红, 基于短语的蒙汉统计机器翻译系统的设计与实现,内蒙师范大学学报(自然科学汉文版), 2011, 40(1): 91 - 94.

11. 百顺, 基于派生文法的日 - 蒙动词短语机器翻译研究, 中文信息学报, 2008, 22(2): 47 - 54.

12. EHARA Terumasa, et al. " Mongolian to Japanese machine translation system. // Proceedings of second international symposium on information and language processing, 2007, 27 - 33.

13. Nagao M. A framwork of a mechanical translation between Japanese and English by analogy principle. In: A. Elithorn and R. Banerji(eds.) artificial and intelligence. NATO publications. 1984.

14. Philipp Koehn, Statistical Machine Translation,UK. Cambridge University,2011.

15. 塚田元, 渡辺太郎 et, al, 統計的機械翻訳,NTT Commincation Technology Instutite,http://www.ntt. co. jp/journal/.

16. P. Koehn,H Hoang,A Birch, et al. Moses: open source toolkit for staistical machine translation // Proceed. Of ACL,2007:177 - 180.

17. Tillmann. C. and Zhang, T.: A Localized Prediction Model for Statistical Machine Translation // Proc. 43rd Annaual Meeting of the Association for Computational Linguiatics, AAssociation for Computational Linguistics, 2005, 557 - 564.

18. Jun. Ye, Cosine Similarity measures for intuitionistic fuzzy sets and their Applications. Mathmatical andComputer Modeling, 2011, Vol. 53,91 - 97.

[责任编辑:奥其]