第 33 卷 第 4 期 《新疆师范大学学报》(自然科学版) Vol.33,No.4
2014 年 12 月 Journal of Xinjiang Normal University Dec.2014
(Natural Sciences Edition)

丝绸之路经济带相似语言信息横向处理通信技术的研究

伊·达瓦[1,2], 米尔阿迪力江·麦麦提[1,2*]
(1.新疆大学 信息科学与工程学院,新疆 乌鲁木齐 830046;2.新疆大学 多语言技术重点实验室,新疆 乌鲁木齐 830046)

摘 要: 丝绸之路经济带为丝路沿线各国及地区带来的发展机遇。丝路沿线不同语言信息的交换及处理技术的提升将面临新的机遇与挑战。以蒙古文为实例,研究了丝路带相似多语言信息交叉处理通信问题,研讨了基于词典及语言学知识和基于统计的机器翻译(SMT)两种方式的蒙文多语种文本的自动转写或翻译方法。试验结果显示,SMT 方法对于相似语言转写的效果相比于语言学以及不同语系不同语族语言的翻译效果近提升一倍。

关键词: 丝绸之路经济带; 相似语言交叉处理通信; 蒙文多文种转写;统计机器翻译;Moses;余弦相似度
中图分类号: TP391 **文献标识码:** A **文章编号:** 1008-9659(2014)04-0066-09
DOI:10.14100/j.cnki.1008-9659.2014.04.013

丝绸之路横跨亚欧大陆,绵延 7000 多公里,总人口近 30 亿。沿线国家多、居住民族多、语言文字多、文化相同点多。现阶段丝绸之路经济带境外建设重点在中亚。中国-中亚沿线,新疆是重要的交通枢纽中心、商贸物流中心、金融中心、文化科技中心、医疗服务中心,新疆已成为丝绸之路经济带上的核心区。新疆将利用丝绸之路经济带核心区建设,推动跨境电子商务、金融等方面的发展,打造网上丝绸之路,利用互联网丝绸之路经济带建设注入活力。

中国-中亚丝路沿线多民族文化有着许多共同点。民俗,宗教以及使用语言环境较为相近。图 1(a,b)为丝路沿线阿勒泰语系多语言不同文字电子化实例文本。其中图 1a 为维哈柯不同民族现用文字语言实例,而图 1b 为蒙古语不同地区现用文字语言实例。其特点是,这类语言的语句法关系,构词结构以及语序上有较大的共性。图 2 为先行研究分析结果[1]。如果把这类语言文字用同一种文字符号(用拉丁文字)转写,其相关性是较大的。所以,相似的多语言信息交叉处理通信技术的开发利用是当务之急。

图 1a 维哈柯不同民族现用文字语言实例

[收稿日期] 2014-10-25
[基金项目] 国家自然科技基金(61163030);自治区科技支疆项目 (编号:201291116)资助。
[作者简介] 伊·达瓦(1956-),男(蒙古族),新疆塔城人,博士,主要从事计算语言学,人工智能学,自然语言信息处理及机器翻译方向研究。
*[通讯作者] 米尔阿迪力江·麦麦提(1989-),男(维吾尔族),新疆喀什人,研究生,主要从事多语言信息处理方向研究。

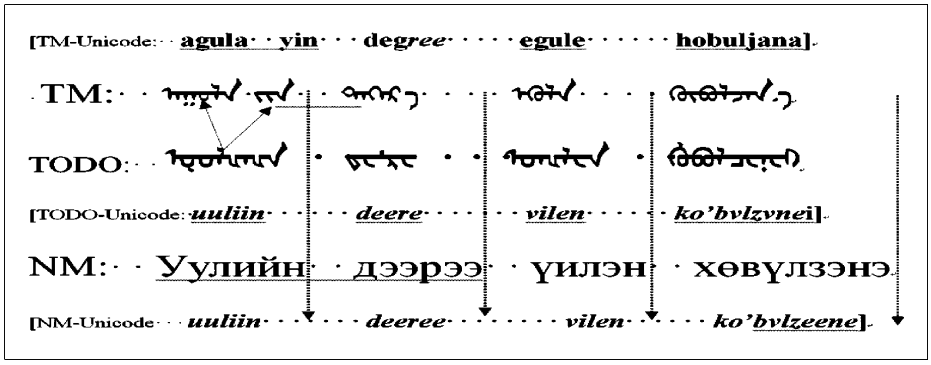


图 1b　蒙古语不同地区现用文字语言实例

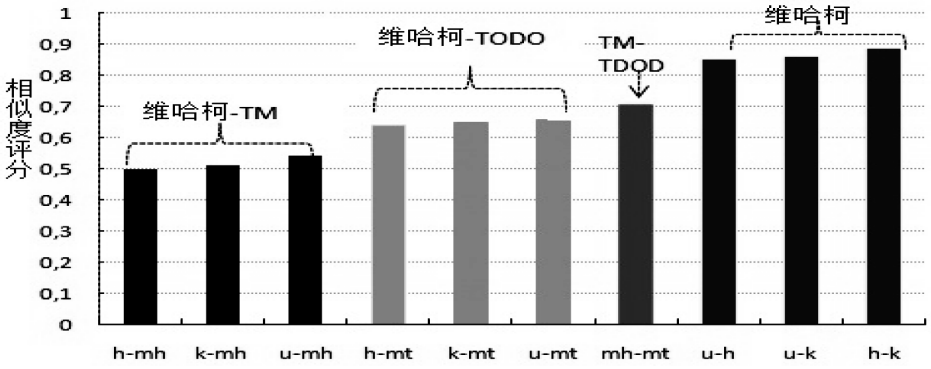


图 2　维哈柯及蒙古语多地文字语言相关性分析结果

文章以蒙古文为实例，尝试研究丝绸之路经济带相似多语言信息交叉处理通信问题。

1　蒙文信息处理现状分析

1.1　蒙文使用情况

蒙古语属阿尔泰语系蒙古语族，属于黏着语。目前使用蒙文地区分布在中国蒙古族聚居区、蒙古国和俄罗斯联邦部分地区。蒙古国以现行新蒙文字（NM）为官方语言，中国境内内蒙地区以及新疆等地区除了使用传统蒙文 TM（traditional Mongolian）文字之外，新疆及俄国的卡尔梅克地区还是用托忒文字（TODO）[2]。图 3 为用这三种文字编写同一句子的电子样本。以上三种蒙文文字，不仅构词和造句规律有区别，而且发音规律也不一致。目前，不同文字电子文本的阅读，通信以及出版还是通过人工转写方式进行。为了减轻人工转写负担，改善编辑出版以及不同文字地区间的通信环境，我们提出了基于计算机技术的转写翻译方法。

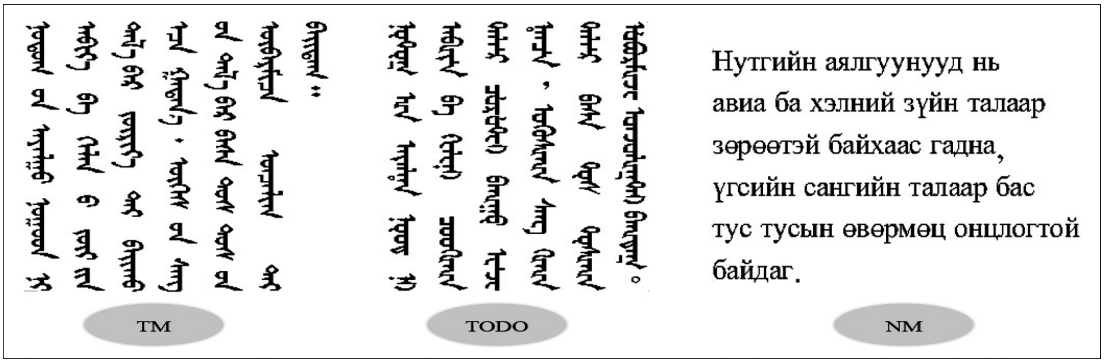


图 3　蒙文常用电子文本样本

1.2　蒙文信息处理现状及存在问题

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，蒙文信息化处理的难度更大。主要有以下几个主要原因：①托忒文、锡伯文、满文名义字符、变形显现字符及控制符使用规则国家标准修订结果尚未出台；②微软提供 Win7/8 及 Office2007 版本输入法环境只提

供了 TM 文字的输入支持,而尚未实现托忒文输入环境的支持[3,4];②GB/T 26226 标准蒙古文使用规则尚未出台;③托忒文拉丁字母转写标准尚未制定,托忒文现用键盘布局与常用规则不吻合;④现有蒙文处理商业软件种类多,标准不统一,相互不兼容;新疆地区使用蒙文网络手机通信环境尚未开发利用。⑤现用蒙文信息处理软件智能化,自动化程度低,技术落后;⑥蒙文自然语言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乏,新疆地区使用蒙文类语言文字数字化应用软件的研发工作几乎处于空白状态。

1.3 蒙文多文种的互换问题

虽然蒙文多文种语言学语法规则,手写方式以及语序相近,但是使用字符形式不同,大多数字符的 Unicode 代码制定也有所不同。比如图 4 所示,蒙文一词/uula "山"/,在 TM,TODO 及 NM 中都发[uual]的音,但各自使用字母符号,录入代码以及字符串的长短均不同。再者,前图 1(b),发现图中用空格区分的句对关系也不一致的。在 TODO 和 NM 句子中功能词(*带色部分*)常连接于前字符串,而 TM 文种功能词必须用空格区分录入,即这三中文中词长,句长也不同。假如,在键盘录入一个 TM 文句子/agvla yin degere/,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,更不能得到 TODO 文短语/ᠤᠤᠯᠠ ᠤᠨ ᠳᠡᠭᠡᠷᠡ/。甚至代码串/agvla yin degere/在用计算机语音合成技术实施发音时,发不出/uula in deere/的语音流。

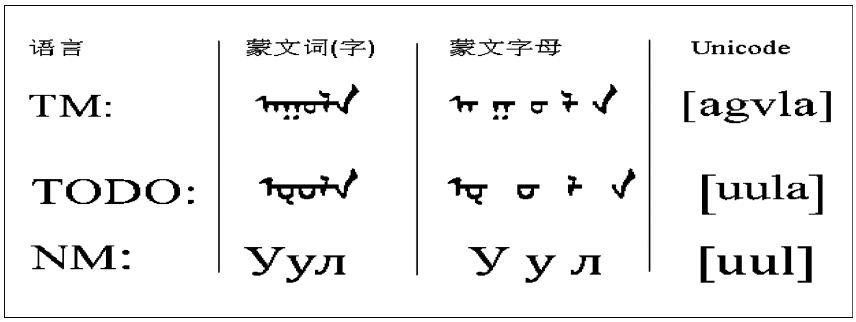


图 4 蒙文三种文字词/uula 山/的对齐显现形式

总之,现用蒙文多文本之间的转换及通信,既不能用字符单位一一对换,又不能用一词或空格区分的字符串形式相互代替。这类相似语言信息交叉处理问题只能通过智能化的自然语言处理前沿技术来实现[5]。

1.4 相关研究状况

目前,有关蒙文多文种文本转换方面的研究并不多。日本筑波大学 Ishikawa 研究组研讨了基于语法规则的 TM 文和 NM 文双向转写方法[6]。该文报告只涉及 TM 和 NM 文处理,没有讨论与 TODO 文的转写处理。蒙古国 Y. Namsurai's 研究组报道了基于语法规则及用少量 NM 文的特定小说数据,尝试 NM 文到 TM 文的转写实验[7]。近来,基于统计机器翻译的汉-蒙(TM)机器翻译的研究[8,9]或日-蒙(TM)机器翻译的研究也有不少的研究报告[10,11],但只限 TM 文本。而 TM-TODO 或更多语言文本间的机器转写或机器翻译的相关研究较少见。我们在先行研究[1,5]中也尝试了基于数据库及基于语言学规则的蒙文多文种的转换处理实验。为此,该文在先行研究的基础上,提出了基于短语统计机器翻译技术的蒙文多文种转写处理新方案。

2 统计机器翻译技术

目前基于计算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的统计机器翻译方法 SMT(Statistical MT)等三大技术模式。

RBMT 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来说,一一描述其规则是不现实的。

EBMT 模式也是一种基于语料库的方法。它通过预设实例句对比原理进行翻译[12]。对于活生生的自然语言句子成分,跟踪设立实例模板也是较难的事。

SMT 模式把机器翻译看成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。它将原语言句

子到目标语言句子的翻译做为一个概率问题进行处理,因为任何一个目标语言句子都有可能是任何一个原语言句子的译文,只是概率不同而已,其翻译即为原文通过模型转换为译文的解码过程,互译时需要找到概率最大的句子。因此 SMT 不需要任何语言学知识、语法规则及实例模板等,只要平行语料规模足够大,一般都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, SMT 技术避免复杂的语法调序处理就能获得高效译文。

针对丝绸之路经济带多个相似语言(语序相同、语法结构一致的特点)的交叉处理通信问题,SMT 方法是否优越,本文以蒙古语多文种为例考查对象,选用基于短语的统计机器翻译方法,引用通用 Moses 解码器[13]尝试相似语言转写处理实现问题。

2.1　翻译系统的构造及工作原理

基于统计翻译的实验系统主要由四个模块组成:①训练语料库: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;②基于对齐语料的翻译模型,如 TM Model (Translation Model);③目标语言统计语言模型 LM (Language Model); ④解码器(Decoder) (见图 5)。工作过程简述如下:首先,研究开发蒙文多语种句子平行文本语料。本研究协助于蒙文媒体出版发行单位,邀请蒙文专家,人工建立了 6 万个句子的 NM-TM-TODO 平行语料(数据实例见图 6),并参考 TM 文功能词(见表 1)信息及双语句长对互译语言实施词对齐处理,生成句对语料。

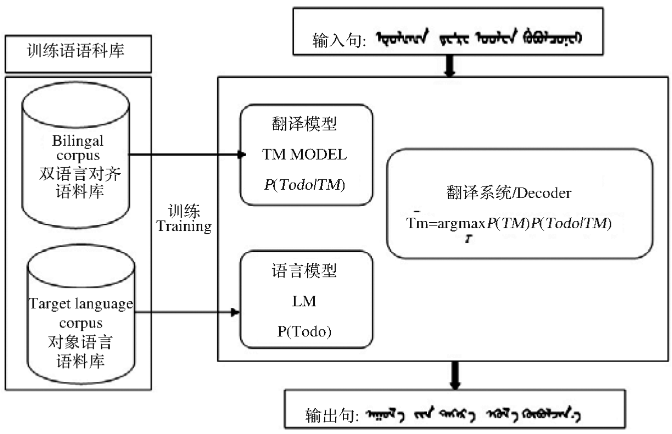


图 5　基于短语的统计翻译的蒙文多文种转写系统

其次,建立目标语大规模文本语料。再次,利用双语对齐语料并借助 Moses 软件,训练从原语言 f 到目标语言 e 的统计翻译模型: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 解码器对原语言 TODO 文句子(如/[illegible]/)实施翻译。

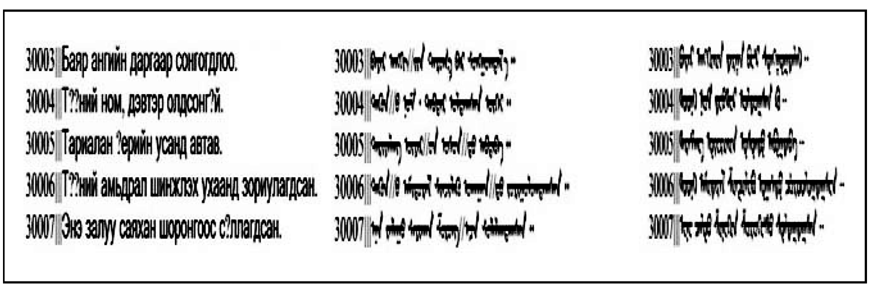


图 6　蒙文多文种句对语料样本

表 1　TM 文功能词列表

a function words set of TM text system
[illegible]

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 e_k (k=1..K),然后在目标语译句子中选拔匹配概率最高的短语片,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由(1)式给出:

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

2.2 句子短语片的排列(调序)

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,如图 7 所示),实施语言 f 和语言 e 的互译。虽然平行语料中双语句子是对齐的,但是句子中短语片之间往往不是按原语言 f 和目标语言 e 词出现的顺序对齐的。因此,系统先对原语言中每个短语片进行互译完毕后,再把翻译短语片按目标语言语法规律重新排序组合而生成目标语言句子(短语片串)。这种作业在统计机器翻译系统中叫做译文短语的调序。

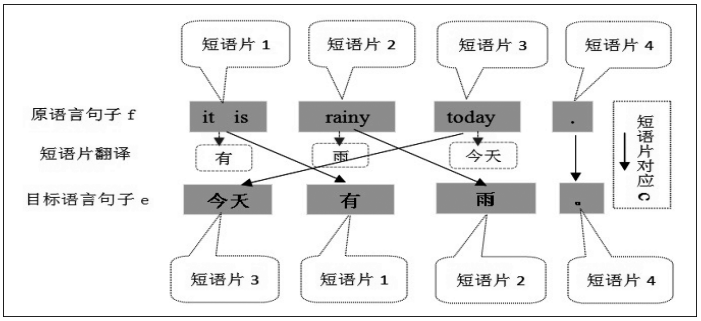


图 7 在短语对齐机器翻译中短语调序方法

在统计机器翻译过程中,互译语言之间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重要且复杂的研究内容。基于短语的统计翻译系统中,考虑短语片出现顺序的翻译模型用(2)式表示:

$$P(f,c \mid e) = P(c \mid e) P(f \mid c,e) \approx P(c_1^I \mid e) \prod_{i=1}^{I} P(\bar{f}_{c_i} \mid \bar{e}_i) \tag{2}$$

其中,$\bar{f},\bar{e}$ 分别是原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,$\bar{e}_i$ 对应原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 所示的句对中,原语言各短语片的翻译顺序为/有 雨 今天 。/,经调序处理后,输出目标语言句子为/今天 有 雨 。/。原语言句子中短语片 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。常用调序算法为 LBO (Lexicalized Block Orientation)模型[14],表示为(3)式:

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

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*。两个短语片位置不相邻,而且顺序与原语言短语片位置偏离时选 *discontinuous* 类进行调序。LBO 模型利用上述 3 类方法进行调序,近似于(2)式的各短语片调序的概率之和,表示为(4)式:

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

2.3 蒙-蒙对齐及调序

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,如图 1(b)中往往存在一对多的情况。其主要原因是 TM 文句子中,功能词不能与前词项像 TODO 和 NM 词那样语法连接。因此本研究采取以下方法进行对齐处理:首先从 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以空格区分对齐处理,否则对 TM 句子端(如图 8)强制功能词与前词项用特殊符号(比如//)做无空格连接,使得原语言句子语序尽

可能地调整为与目标语言一致。在此调序模型(4)式按 *monotone* 类进行调序,则句对按空格顺序 $c_{i+1} = c_i + 1$ 得到排序。

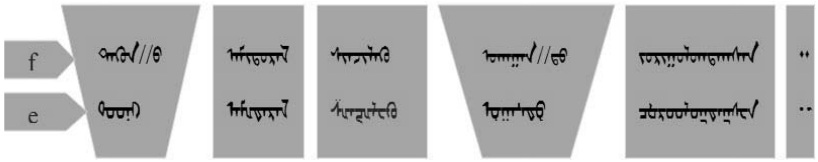


图 8　f 和 e 双语句子以功能词强迫空格对齐处理

3　系统测试实验

3.1　实验条件

研究开发了以人工选择录入的 NM、TM 及 TODO 三种文字文本,共录入了 6 万条句子平行语料,供评测实验使用。对于短语表的生成及调序模型的学习用相同语料。考虑到蒙文多文种语言的语序相同,语法相似及试验语料有限等因素,本次评测设置了三组实验。

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

第一组,用 5 万词蒙文多文种标注电子词典[15]对于原语言句子空格区分词项;检索目标语言匹配词项生成目标语言句子,再引用 BLEU 值和余弦相似法考察互译句对的相似性;第二组,以标准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;第三组,对于 TM 句子进行功能词与前词项强制连接。

3.2　实验 1:基于词典的转写方法试验

3.2.1　实验方法

利用图 9 所示的 5 万词条的中-蒙文多文种词标注词典,做两次 TM　TODO 方向句子转写测试:① 对于原语言 TM 句子中的每个用空格区分词,经词典检索匹配 TODO 词,再与 TM 句子同序排列输出 TODO 句子。② 对于原语言 TM 句子中的每个用空格区分词,经词典,引用式(7)余弦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。

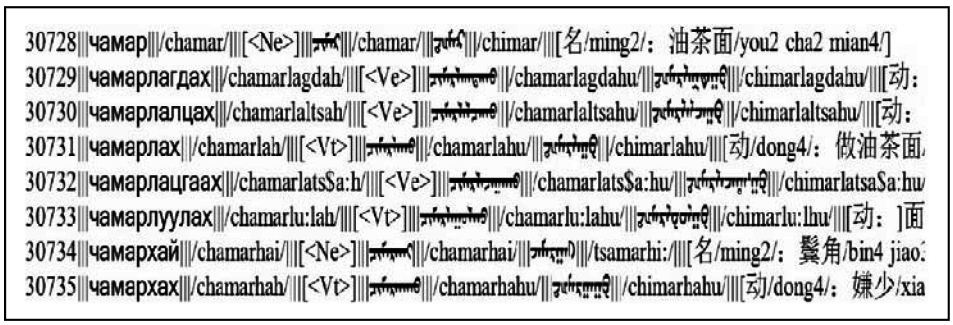


图 9　蒙-汉对齐电子词典样本

3.2.2 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 cossine$\theta = 1$,

(θ=0°)时,两个向量 A 和 B 相同,即 A 和 B 完全相似;当 cosineθ=0,(θ=90°)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[16]。

$$\begin{cases} A = a_1, a_2, \cdots a_n \\ B = b_1, b_2, \cdots b_m \end{cases} \tag{6}$$

$$sim = \cos(\theta) = \frac{A \cdot B}{\| A \| \ \| B \|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{m} (B_i)^2}} \tag{7}$$

3.2.3 BLEU 尺度:互译文自动评估常用尺度,依据 N-gram 匹配率,见(8)式:

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

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

C 为 MT 译句长度,r 为参考译句中最近译句长度[17]。

3.2.4 实验 1 结果

利用以上两种方法,首先对原语言 TM 测试集 300 句子转写输出 TODO 文 300 句子, 然后算出其结果与 TODO 文 250 个测试句的 BLUE 值和句子相似度 *sim*。其评估结果见表 3。

表 3 基于词典测试结果

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

3.3 实验 2:基于短语的统计翻译转写

本次实验实施 *TM↔Todo* 双向翻译。实验数据与表 2 相同。评测用 BLEU 值和句子相似度 *sim* 分别进行评估。在基线测试中 TM 文功能词与前词项不做强制连接,而提案方法用 TM 功能词与前词做强制连接后的句对语料,分别进行评估。

为了增大目标语言句子的变异性,在获取 N-best 翻译结果时,只将得分最高候补句子选入 N-best 中。在目标语 N-gram 语言模型的学习中,除用表 2 所示语料以外,还增加了不同领域的文本。引用开原工具 SRILM 加入 Kneser-Ney,平衡参数进行 1-5 元语言模型训练。其中 3-元模型训练设定为://ngram-count - text train.tex - lm lm - interpolate - kndiscount1 - kndiscount2 - kndiacount3-order 3,解码器相关参数见表 4。各参数权重学习用开发集实施 Minimum Error-Rate Training 方法。

表 4 评测参数设定

项目	条件
解码器	Moses (10/03/10)
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 5　基于短语机器翻译评测结果

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

表 5 中给出基于短语的统计机器翻译实验 2 的评测结果。比较实验结果表 3 和表 5 可以看到，提案（proposed）统计机器翻译方法的最好 BLEU 值（TODO　TM）58.03%比基线（baseline）方法的 BLEU 值 54.87%高出 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 比基于词典的相似度 0.68 约高出 0.2。译句与目标语测试集的相似性大幅度提高。另外我们还发现，该提案方法评测结果明显高出先行汉-民机器翻译研究目前最好的翻译成果 29.86%近两倍[8]。这充分证明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及优越性。

4　结论

通过提案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相似语言文本信息的转换处理极为有效。对于蒙文多文种文本的转写处理，着重 TM 文功能词，强制合并于前词而实现对 TODO（NM）文句子的词对齐关系，提高短语统计翻译的译文质量。同时，实验结果也显示，基于语言学知识及依赖词典的互换方式处理过程复杂，互译质量明显低于统计翻译效果。这完全反映了统计机器翻译对于语言学及语序差异性较小语言互译的优势。通过本次实验也体会到，这种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便利快捷的方法。

现有多语种对齐语料有限，且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，对本次实验结果有一定的影响。扩大多文种平行语料资原规模，改善现有语言资原质量以及提升系统性能，是今后本研究工作的重点。

参考文献：

[1] 达瓦·伊德木草，木合亚提. 维哈柯及蒙语多文种语言相似性考察研究[J].中文信息学报，2011，27(6)：180-186.

[2] D.Tserenpil, R.Kullmann, Mongolian Grammar. School of Mongolian Language and Culture, National University of Mongolia and Institute of Language and Literature[C]. Academy of Science, Mongolia，2005.

[3] 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.

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

[5] Idomucogiin Dawa, Satoshi Nakamura. A Study on Cross Transformation of Mongolian Family Language[J].Journl of Natural Language Processing, J-STAGE, 2008，15(5)：3-21.

[6] T. Ishikawa，. et al. A Bidirectional Translation Method for the Traditional and Modern MongolianScripts[C]. Proceeding of the Eleventh Annual Meeting of The Association for Natural Language Processing, 2005：360-363.

[7] Y.Namsurai. et al. The database Structure for BI-Directional Textual Transformation Between TwoMongolian Scripts[C]. Proceeding ICEIC，2006：265-268.

[8] 王斯日古楞，等，汉蒙统计机器翻译中调序方法研究[J]. 中文信息学报，2011，25(4)：88-92.

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

[10] EHARA Terumasa, et al. Mongolian to Japanese machine translation system [C]. Proceedings of second international symposium on information and language processing, 2007：27-33.

[11] 陈雷，李森等.有限语料汉蒙统计机器翻译调序方法研究[J].中文信息学报，2013，27(5)：198-203.

[12] 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[M]. NATO publications, 1984.

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

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

[15] I.Dawa, U.Aishan, A.Kahaerjiang, Y.Turgen. Creating and Analysis of a POS Tag Chinese-MongolianMulti-Version Dictionary[J]. International Journal of Computer Mathematics, 2014,3(16):123-127.

[16] Jun. Ye. Cosine Similarity measures for intuitionistic fuzzy sets and their Applications[J]. Mathmatical and Computer Modeling, 2011,53: 91-97.

[17] Philipp Koehn. Statistical Machine Translation[M]. UK. Cambridge University,2011.

An Investigation Study on Cross Transformation Between the Similar Languages Along the Silk Road Economic Belt

I · Dawa[1,2], Mieradilijiang · MAIMAITI[1,2]

1. *College of information science &engineering, Urumqi, Xinjiang*, 830046, *China*;

2. *Xinjiang Laboratory of Multi-language Information Technology, Xinjiang University, Urumqi, Xinjiang*, 830046, *China*)

Abstract: With the voice of developing the Silk Road economic belt (SREB), the communication between the multiple languages and their processing along SREB become a practical concern. Focus on Mongolian multi-graphic texts, widely used Mongolia, China and Russia today, in this research, we investigate two kinds of approaches between their converting. One is interested in the linguistic knowledge applied a dictionary, and other is focus on the phrase-based statistical machine translation (SMT). Results from the experiments demonstrate that the approach SMT is more convenient than the linguistic knowledge way. Further, we confirmed that SMT result for the similar language transformation is better than that for the different language family such as Chinese -Mongolian STM reported in previous research.

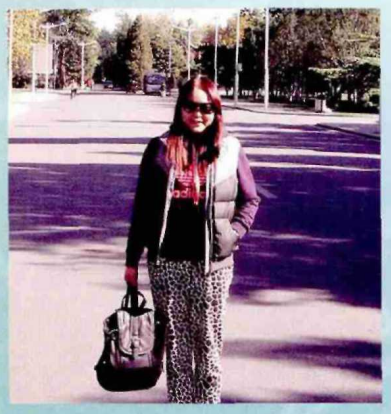
Key words: Silk Road Economic Belt; Similar Language Cross Processing; Mongolian text converting; phrase -based SMT; Moses; Cosine similarity

本期特约稿件作者

AUTHORS FOR SPECIAL ARTICLES

伊·达瓦，新疆大学信息科学与工程学院教授，博士。自治区多语言信息技术重点实验室学术带头人，新疆大学信息科学与工程学院以及自治区经济与信息化委员会委托博士生导师；武汉理工大学计算机技术应用学院委托博士生导师。现主持国家级科研项目2项；省部级科研项目4项；国内外发表论文共72篇；在核心期刊发表学术论文15篇。国家自然科学基金委特聘评审专家，国家计算语言学会学报评审专家，亚太地区信息技术与计算机应用学国际会议评审专家。曾被聘用到日本国家信息技术通信研究机构、国际通信研究所的高级研究员；同时担任早稻田大学国际通信研究科客座教授、博士生导师。

李丙春，新疆喀什大学信息工程技术系主任，工学硕士学位，教授，历任喀什师范学院网络中心主任。现从事图像处理、模式识别和智能计算等方面的研究，在核心期刊发表专业学术论文多篇。参与国家自然科学基金项目1项和新疆维吾尔自治区高校科研计划重点项目2项。主持完成喀什大学精品课程1门，主编教材1部，参编教材1部。获喀什大学2014年“教学名师”称号。



鲍晓玲，新疆师范大学体育学院副教授，田径国家级裁判员。主要从事体育锻炼与身心健康方面的研究。主持自治区课题1项，在核心期刊发表论文5篇，在省级刊物发表论文10余篇，获第十二届中学生科报会三等奖。