

融合深度学习特征的汉维短语表过滤研究

朱顺乐

(浙江海洋大学, 浙江 舟山 316000)

摘要:汉维机器翻译面临着汉维语言构词、语序差异性大,短语表冗余、不合理信息较多,双语资源匮乏以及相应形态分析工具性能欠佳等挑战,严重影响了汉维机器翻译译文质量。针对汉维短语表中出现较多的不合理短语对,影响翻译性能及解码效率这一问题,提出一种融合汉维短语对循环神经网络特征和汉维短语对上下文特征等深度学习特征,以及汉维短语对平均词共现特征这一浅层特征的汉维短语表过滤模型。该模型基于短语对循环神经网络特征、上下文特征以及平均词共现特征,并将各个特征概率及训练实例输入到基于朴素贝叶斯分类器的短语表过滤模型进行训练。该模型结合了汉维候选短语之间更为丰富的语义及上下文信息。实验结果表明,提出的短语表过滤方法能够有效地去除汉维短语表中的不合理短语,汉维机器翻译性能及其解码效率都有所提高。

关键词:循环神经网络;贝叶斯定理;非连续元;短语表过滤;汉维翻译

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2018)07-0149-06

doi:10.3969/j.issn.1673-629X.2018.07.032

Research on Chinese-Uyghur Phrase Table Filtering Integrating Deep Learning Features

ZHU Shun-le

(Zhejiang Ocean University, Zhoushan 316000, China)

Abstract: Chinese-Uyghur machine translation is faced with challenges such as difference of word formation and word order between Chinese and Uyghur, phrase table redundancy, unreasonable phrase pairs, lacking of bilingual resources and poor performance of corresponding morphological analysis tools, which seriously affect the performance of Chinese-Uyghur machine translation model. To solve these problems in Chinese-Uyghur phrase table that many unreasonable phrase pairs exist and affect the performance and productivity of translation model, we propose a Chinese-Uyghur phrase table filtering model integrating deep learning features like recurrent neural network feature and context feature of Chinese-Uyghur phrase pair and shallow feature like average co-occurrence feature. The model is on the basis of phrases for circulation neural network feature, context feature, and the average word co-occurrence feature, and the characteristics of probability and examples of training are input to phrases list filtering model based on Naive Bayesian classifier for training. This model combines the richer semantic and contextual information between the candidate phrases of Chinese-Uyghur. Experiment shows that the proposed phrase table filtering method can effectively eliminate the unreasonable phrases in the phrase table of Chinese-Uyghur and improve the translation performance and decoding efficiency of Chinese-Uyghur translation machine.

Key words: recurrent neural network; Naïve Bayes; skip-gram; phrase table filtering; Chinese-Uyghur translation

1 概述

随着经济全球化的不断深入,国家与国家之间、民族与民族之间交流时的语言障碍突显,已成为经济发展、文化交流的不利因素。机器翻译技术的发展为缓解这一障碍提供了契机。统计机器翻译(statistical machine translation, SMT)是目前学术界研究的主流方法。它是非限定领域机器翻译中性能较佳的一种方

法。其基本思想,是通过对大量的平行语料进行统计分析,构建翻译模型(translation model, TM),对目标语言单语语料进行统计建模,构建语言模型(language model, LM),进而使用上述模型对输入源语言句子进行翻译。

统计机器翻译模型又分为基于词的翻译模型、基于短语的翻译模型以及基于句法的翻译模型三类。其

收稿日期:2017-08-03

修回日期:2017-12-21

网络出版时间:2018-03-07

基金项目:浙江省自然科学基金资助项目(LY16F020014);浙江省自然科学基金青年科学基金项目(LQ16A010003)

作者简介:朱顺乐(1977-),男,讲师,研究方向为自然语言处理、机器翻译、农业信息化。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20180307.1417.020.html>

中,基于短语的翻译模型既在翻译过程中考虑到了局部上下文信息,又不需要句法标注语料,并且能取得较好的翻译效果,因而广受学术界与工业界的青睐。

$$\tilde{e} = \arg \max_{e \in e^+} (e | f) \quad (1)$$

其中, f 表示源语言句子(汉语); e 表示目标语言句子(维吾尔语); \tilde{e} 表示最佳翻译。

汉维翻译模型训练架构如图 1 所示。

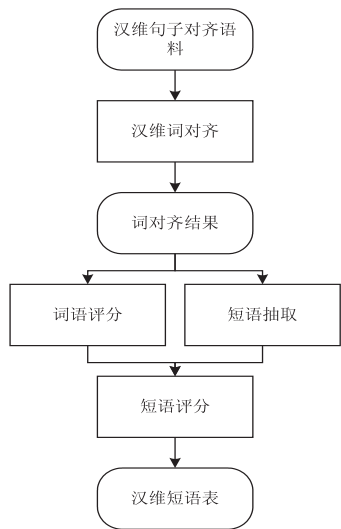


图 1 汉维翻译模型训练架构

作为基于短语机器翻译模型框架的核心部分,翻译模型提供短语表、调序规则表等重要知识。短语表中包含双语短语的互译信息,其质量直接影响机器翻译模型的性能。然而,以下两个因素会对短语表的质量以及后期解码效率产生影响。(1)短语表抽取位于统计机器翻译框架的中间环节,前期的词对齐阶段产生的错误会延续到短语表生成阶段;(2)统计机器翻译模型性能很大程度上依赖于双语句子平行语料。目前,日益丰富的网络资源使得大规模语言资源的获取成为可能,然而大规模语料使得双语短语表规模呈指数级增长,从而减了解码速度。因此,对短语表中的噪音短语进行过滤,增大了解码阶段解码器检索到更为准确的翻译片段的概率;非法短语对的过滤可以减小短语表的规模,一定程度上提升了解码效率。

针对短语表过滤这一任务,国内外学者进行了一些研究。Nishino 等提出一种基于子模函数最大化的短语表过滤方法,采用贪心的启发式算法策略实现^[1];Wang 等提出一种面向短语表过滤的相对熵模型,并用其衡量用小概率的翻译事件推导出短语对表示翻译事件的概率^[2];Azadi 等使用主题模型进行短语表的剪枝^[3];Zens 等首先比较了多种短语表过滤方法,并提出了基于语音理论的短语表过滤框架^[4];Torr 提出了一种基于句法的短语表过滤模型,该模型依赖于句法分析的语料数据。

针对汉维机器翻译的相关研究开展较晚,前期的研究主要集中在语言的分析^[6-8]、语料库建设^[9]、命名实体识别^[10-12]以及翻译系统构建^[13-16]等方面。对于短语表的过滤及其相关工作的研究较少。

前期的研究工作并没有考虑短语的上下文信息以及双语的语义关系,即使有基于句法的模型,也要依赖于大规模的句法标注语料。文中提出一种新颖的汉维短语表过滤方法,将短语表的过滤看作分类问题:基于朴素贝叶斯(Naïve Bayes, NB)模型,融合了短语对循环神经网络(recurrent neural network, RNN)特征、上下文特征等深度学习特征,以及平均词共现特征等浅层特征,获得汉维短语对是否保留的概率值,并通过实验对其进行验证。

2 短语表生成

基于短语翻译模型中的短语表依赖于词对齐阶段产生的对齐文件以及汉维平行语料构建。因此短语表的创建可分为两个阶段:词对齐矩阵生成和短语表抽取。

2.1 词对齐矩阵生成

统计机器翻译中的词对齐,是基于统计学习的相关算法,从大规模的双语句子平行语料中自动获取词语共现等的过程。使用较多的词对齐算法包括 IBM Model 1-5^[16-17]以及基于(hidden Markov model, HMM)的词对齐模型^[18]。Och 基于上述模型,设计开发了广泛使用的词对齐开源工具 GIZA++^[19]。词对齐矩阵^[20-21]即是根据词对齐结果生成的,见图 2。

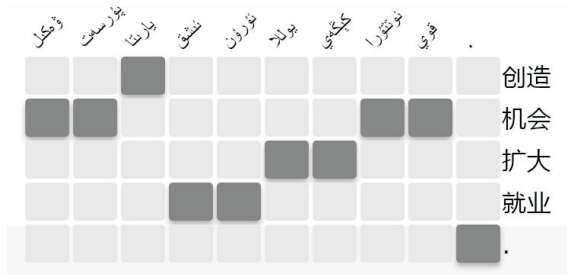


图 2 汉维机器翻译词对齐矩阵

2.2 短语抽取

汉维短语抽取是短语表创建的基础。基于词对齐矩阵获取汉维双语短语对的方法如下:若与矩形所在行范围内的汉语词对齐的维吾尔语词也在当前子矩形内,提取对齐矩阵中所有以对齐点为顶点矩形区域所表示的汉维短语对。其核心思想,即是首先穷举汉语句子中所有可能的短语,根据词对齐矩阵,检索对应维吾尔语句子中的短语。抽取的部分汉维短语表如图 3 所示。

上述过程抽取到的只是候选短语。在添加至短语表之前,还应对其进行校验。校验遵循的原则有两个:

"绿色 通道"		«	2	1			2.718	0.0276761	0.5	0.000115966	1		ويچىشل بول		
"绿色 通道"		1.99879	0.025641			ويچىشل بول	-06	0.5	0.178357	2.718			39	2	
"绿色 长城"		3.55212	0.5			ويچىشل سېيىل بىرپا	-06	0.5	0.00241937	2.718			2	2	
"绿色 长城"		3.55212	0.5			ويچىشل سېيىل بىرپا	-06	0.5	0.000251112	2.718			2	2	
"绿色 长城"		3.08247	0.5			ويچىشل سېيىل بىرپا	-06	1	0.000243095	2.718			2	1	
"网络"		»		0.000362188	6.95319e-07	1	0.158621	2.718			2761	1			
"网络 人物"		»	1.48997	0.166667			هەنپەنخىس	-07	1	0.0653856	2.718			6	1
"网络 人物 候选人"		»	6.76646	0.5			هەنپەنخىس نامزات ئالالا	-10	1	2.71877e-05	2.718			2	1
"网络 人物 候选人"		»	6.26356	0.5			هەنپەنخىس نامزات ئالالا	-08	0.5	0.060781	2.718			2	2
"网络 人物 候选人"		»	6.26356	0.5			هەنپەنخىس نامزات ئالالا	-08	0.5	4.31971e-05	2.718			2	2
"老"		4.29527	0.00244499				ەكونا	-05	1	0.20712	2.718			409	1
"老 北京"		1.84759	0.0434783				ەكونا	-05	1	0.190167	2.718			23	1
"老 北京 民俗"		5.67209	0.0434783				ەكونا	-10	1	0.190167	2.718			23	1
"老 北京 民俗 庙会"		3.93227	0.05				بۇتخانا سەيلە	-14	1	0.0438774	2.718			20	1
"老 北京 民俗 庙会" 的		7.61125	0.05				بۇتخانا سەيلە	-15	1	0.0438774	2.718			20	1
"老 北京 民俗 庙会" 的 历史		4.13003	0.166667				بۇتخانا سەيلە	-15	1	0.0321876	2.718			6	1

图3 汉维短语表(局部)

(1) 候选汉语端的单词在汉语句子中的位置必须连续;

(2) 候选汉维短语必须与汉维词对齐矩阵相容。即汉语短语中的词 $c_i^j(j \leq j' \leq j+m)$ 或者对空, 或者对齐到维吾尔语短语 $e_i(i+n)$ 中的词 e_i' , 反之亦然。

统计词对齐模型基于大规模的平行语料。然而, 由于平行语料规模的局限性以及汉语、维吾尔语的差异性, 汉维词对齐过程中会出现数据稀疏问题, 影响了词对齐的准确性, 进而导致汉维短语抽取过程中出现偏差, 影响后续的翻译模型训练以及机器翻译系统的解码效率。

3 特征描述

为了对汉维短语表进行过滤, 从双语短语对循环神经网络特征(RNN)、上下文特征(BIT)以及短语对中平均词共现次数(ACC)等特征出发, 分别构建相应的特征函数。

3.1 短语对循环神经网络特征

为了最大限度获取汉维短语表中候选双语短语的对应关系, 以便更好地对短语表进行过滤, 基于RNN^[21-22]获取维吾尔语和汉语短语之间的互译概率。RNN的主要优势在于处理序列数据。与以往的模型不同, 基于RNN处理序列预测问题时, 该序列当前的输出与之前的输出也有关系, 即网络会对前面的信息进行记忆并应用于当前序列输出的计算中; RNN网络中, 隐藏层之间的节点是有连接的, 当前隐藏层的输入不仅包括输入层的内容, 还包括上一时刻隐藏层的输出。

根据短语表过滤这一应用, 文中使用RNN的编码器-解码器架构。基于该网络结构可以同步获得短语表中汉维短语对的对齐及翻译概率值, 将其作为该短语是否保留的重要特征之一。

将双语短语之间的对应概率进行转换, 用以预测汉维短语词之间的对应关系。使用RNN方法预测*i*时刻的词对应概率可形式化地表示如下:

$$p(y_i | y_1, y_2, \cdots, y_{i-1}, x) = g(y_{i-1}, s_i, \mathbf{c}_i) =$$
$$g[y_{i-1}, f(s_{i-1}, y_{i-1}, \mathbf{c}_i), \sum_{j=1}^{T_x} \alpha_{ij} h_j]$$

(2)

其中, s_i 表示 RNN 模型时刻 t 的隐藏状态; 上下文向量 \mathbf{c}_i 依赖于输入短语映射的标记序列, \mathbf{c}_i 可被定义为对标记 h_i 的加权求和, 标记的权值计算如下:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

(3)

3.2 短语上下文特征

统计机器翻译模型训练过程中存在较为严重的数据稀疏问题。造成数据稀疏的原因是复杂的, 即使使用超大规模的语料库也不能获取每个词组成的所有字符串。训练过程中的数据稀疏问题也会对短语表的过滤产生影响。针对该问题, 提出一种缓解数据稀疏的策略, 即基于Skip-gram^[23-24]获取双语短语中的上下文特征, 计算相应的概率值, 并将其作为短语表过滤模型的特征之一。

Skip-gram 是 n-gram 的泛化。与 n-gram 类似, 也是使用 n-gram 的方式对语言建模, 但允许 n-gram 语法中跳过若干词。Skip-gram 可定义如下:

给定句子 S (以单词序列 w_1, w_2, \cdots, w_m 表示, 其中 m 是句子 S 的长度), 定义 k-skip-gram 集合为: $\{w_{i_1}, w_{i_2}, \cdots, w_{i_m} | \sum_{j=1}^m i_j - i_{j-1} < k\}$ 。表示在句子 S 中, 指定一个 skip 值 k , 允许在 S 的 n-gram 语法中跳过 k 个或小于 k 个单词。

文中的当前词设定为词对齐阶段准确率较高的词, 根据该汉维词对, 预测其上下文信息, 进而获得双语上下文信息中存在的有对齐关系词对的对齐概率。将此概率作为最终的双语短语上下文特征。

$$\Pr_{\text{context}}(E | C) = \frac{\sum_{c \in C} \sum_{e \in E} n(c' * e')}{\sum_{c \in C} \sum_{e \in E} n(c \cdot e)}$$

(4)

其中, C 和 E 是根据 Skip-gram 算法得到的对应

位置元素有语义关系的子短语集合; $c' * e'$ 表明两个单词的对齐概率大于某个阈值 t (经验值)。

3.3 短语对平均词共现特征

汉维双语平行语料包含大量的词对应信息。文中提出的短语表过滤模型中第三个重要的特征即是充分利用汉维平行语料中的词共现信息,提取汉维短语之间的对应关系。具体做法如下:根据词对齐阶段统计的汉维词共现信息,计算得到当前汉语短语对中有互译关系词在短语对中所占比例。

$$\text{Score}_{\text{ACC}} = \frac{\sum_i \text{CoNUM}(c_i, e_j)}{\text{Len}_s}, i < \text{Len}_s, j < \text{Len}_t \quad (5)$$

其中, $\text{CoNUM}(c_i, e_j)$ 表示根据汉维词共现信息,短语对中的汉语词 c 和维吾尔语词 e_j 之间存在对应关系; Len_s 表示汉语短语长度; Len_t 表示维吾尔语短语长度。

4 汉维短语表过滤模型

根据上述短语对循环神经网络特征、汉维双语短语上下文特征以及汉维短语对平均词共现特征以及朴素贝叶斯分类模型,构建面向汉维机器翻译的短语表过滤模型。

4.1 朴素贝叶斯分类器

朴素贝叶斯分类模型^[25]是一种基于特征独立假设贝叶斯定律的简单概率分类器。该分类器可以更加精确地描述特征之间潜在的概率关系。朴素贝叶斯模型基于概率推理过程,即各个条件均存在一定概率的不确定性,在仅仅知道其出现概率的情况下,如何完成分类过程。朴素贝叶斯分类模型基于独立假设,即分类假设样本特征之间是相互独立的。

朴素贝叶斯模型依赖精确的概率推理,因此,与其他分类算法相比,其在有监督学习的样例集合上能获得较好的分类效果,广泛应用于文本分类、数据挖掘等领域。

通过对汉维短语表中抽取出的三个特征进行分析,发现三个特征之间不存在直接的相关性,短语对循环神经网络特征依赖于当前短语所在句子的全局信息;短语上下文特征考虑当前短语对中词在大规模单语语料中的语义关系;平均词共现特征仅仅考虑当前短语对中词之间的对齐信息。因此,文中选择朴素贝叶斯模型作为短语对过滤模型的基线算法。

4.2 短语表过滤模型

文中提出的汉维短语表过滤模型主要由以下三部分组成:原始汉维短语表获取;汉维短语对特征抽取;汉维短语对平均词共现特征。

汉维短语表过滤模型的输入为特征向量 f , 输出为类标记 c 。其中,特征向量包括三类特征:汉维短语对循环神经网络特征、汉维短语对上下文特征以及汉维短语对平均词共现特征。文中提出的短语表过滤模型构成的特征向量可以形式化地表示为: $T = \{(f_1, c_1), (f_2, c_2), \dots, (f_n, c_n)\}$, 其中的特征由三元组组成: $\langle f_{\text{RNN}}, f_{\text{BIT}}, f_{\text{ACC}} \rangle$, f_{RNN} 表示汉维短语对循环神经网络特征概率, f_{BIT} 表示汉维短语对上下文特征概率, f_{ACC} 表示汉维短语对平均词共现特征概率。输出标记 c 为 $\{\text{"reserve"}, \text{"delete"}\}$, 其中, "delete" 表示当前汉维短语对为不合理短语对,应从短语表中过滤; "reserve" 表示当前短语对为合法短语对,应保留。

5 实验

5.1 实验设置

5.1.1 实验数据

为了验证提出的短语表过滤模型的有效性,实验使用了三类语料:汉维双语句子平行语料(训练集:300 000,开发集:700,测试集:1 500)、汉维词典(24 万词)以及人工筛选的汉维短语对正反例(正例 1 000 短语对,反例 1 000 短语对)。其中双语句子平行语料主要用于统计机器翻译模型训练及其双语特征抽取;汉维词典用于双语短语特征获取;汉维短语对正反例用于训练短语对过滤模型。

5.1.2 实验装置

汉维机器翻译实验使用开源的机器翻译工具集 Moses,分别在基于短语模型以及基于层次短语模型上进行实验。语言模型选用 SRLM,使用五元语言模型。参数调整使用 MERT 算法^[26]。机器翻译性能打分使用 BLEU^[27]。汉语端分词工具使用 NLPPIR。汉维短语对循环神经网络特征抽取基于开源的工具集 DeepLearning4j 实现。短语对上下文特征抽取,使用 word2vec 工具实现。汉维短语表过滤模型采用自主实现的 naivebayes4j 训练。

5.2 实验过程

首先,对汉语和维吾尔语语料进行全半角转换、分词、Tokenization 操作;其次,采用双语语料获取原始短语表;再次,抽取汉维正反例语料中的汉维短语对循环神经网络特征、汉维短语对上下文特征以及汉维双语短语平均词共现特征,将其作为输入进行短语表过滤模型训练;最后,采用训练得到的模型在不同短语长度限制下进行短语表过滤实验。

5.3 实验结果与分析

分别从对短语表规模、翻译解码效率以及翻译性能的影响进行分析。

5.3.1 对汉维短语表的影响

根据提出的短语表过滤模型,基于短语汉维机器翻译短语表的规模在最大短语长度分别取 7,9,11 时均有较大幅度减小(见表 1)。为了验证文中方法的泛

表 1 对汉维短语表(规则表)规模的影响

	Phrase-based model			Hierarchical phrase-based model		
	len = 7	len = 9	len = 11	len = 7	len = 9	len = 11
Baseline	2 178 300	3 238 290	3 612 812	6 827 430	10 543 700	16 077 201
Ours	1 224 308	1 740 300	1 890 400	4 234 000	9 094 620	12 741 092

5.3.2 对机器翻译效率的影响

从表 1 可以看出,由于大量不合理短语(规则)对被文中提出的模型过滤,短语(规则)表规模有了明显减小。因此,解码的效率也有所提高(见表 2)。

表 2 对汉维翻译解码效率的影响

	Phrase-based model			Hierarchical phrase-based model		
	len = 7	len = 9	len = 11	len = 7	len = 9	len = 11
Baseline	540	692	815	1 129	1 242	1 610
Ours	413	528	617	1 025	1 090	1 500

5.3.3 对模型性能的影响

由于提出的短语表过滤模型一定程度上减少了不合理短语对的数量,过滤后的汉维机器翻译质量总体高于过滤前。对比短语模型和层次短语模型,在规则长度不少于 9 时,层次短语模型翻译质量高于短语翻译模型。其中,最大规则长度为 9 时,基于层次短语的汉维机器翻译模型在过滤后翻译性能达到最优(见表 3)。分析原因,与基于短语的模型相比,层次短语模型中的非终结符有一定的泛化能力及局部调序能力。

表 3 对汉维机器翻译模型性能的影响

	Phrase-based model			Hierarchical phrase-based model		
	len = 7	len = 9	len = 11	len = 7	len = 9	len = 11
Baseline	0.273 3	0.281 0	0.272 7	0.279 8	0.282 0	0.279 1
Ours	0.277 9	0.288 4	0.274 0	0.280 7	0.296 2	0.280 2

6 结束语

由于汉语和维吾尔语在构词及形态上存在较大差异性,模型训练过程中存在较严重的数据稀疏问题,致使汉维词对齐出现偏差;这一偏差又会传递至短语表生成阶段,产生不合理的短语对,最终影响翻译质量机器解码效率。综合考虑汉、维吾尔语言特征及汉维短语表中存在的问题,提出了一种融合深度学习特征的汉维短语表过滤模型,该模型基于短语对循环神经网络特征、上下文特征以及平均词共现特征,并将各个特征概率及训练实例输入到基于朴素贝叶斯分类器的短语表过滤模型进行训练。该模型结合了汉维候选短

语化功能,也在汉维层次短语模型上进行了实验,在最大规则长度分别取上述值时,规则表规模也有所减小。分析原因,提出的方法过滤了大量的不合理短语(规则)对。

之间更为丰富的语义及上下文信息。实验结果表明,该方法有效提升了汉维机器翻译性能,解码效率也有了显著提高。

在下一步的工作中,将在该模型的基础上融入更多的语言学信息,如词性标注、句法标注等,以更大幅度地改善汉维机器翻译质量及其翻译效率。

参考文献:

[1] NISHINO M, SUZUKI J, NAGATA M. Phrase table pruning via submodular function maximization[C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin: Association for Computational Linguistics, 2016:406-411.

[2] WANG Ling, GRACA J, TRANCOSO I, et al. Entropy-based pruning for phrase-based machine translation[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island, Korea: Association for Computational Linguistics, 2012:962-971.

[3] AZADI F, KHADIVI S. Phrase table pruning by modeling the content of phrases[C]//7th international symposium on telecommunications. Tehran, Iran: IEEE, 2014:535-538.

[4] ZENS R, STANTON D, XU Peng. A systematic comparison of phrase table pruning techniques[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island, Korea: Association for Computational Linguistics, 2012:972-983.

[5] TORR J. Syntax-based phrase-table pruning for statistical machine translation[D]. Edinburgh: University of Edinburgh, 2014.

[6] 麦热哈巴·艾力,姜文斌,王志洋,等.维吾尔语词法分析的有向图模型[J].软件学报,2012,23(12):3115-3129.

[7] 米莉万·雪合来提,麦热哈巴·艾力,吐尔根·依布拉克,等.维吾尔语词尾对汉维统计机器翻译影响的研究[J].计算机工程,2014,40(3):224-227.

[8] 米莉万·雪合来提,刘凯,吐尔根·依布拉克.基于维吾尔语词干词缀粒度的汉维机器翻译[J].中文信息学报,2015,29(3):201-206.

[9] 于清,孙浩男,陈永杰.汉维医疗平行语料库构建及特征

- 分析[J]. 新疆大学学报: 自然科学版, 2017, 34(2): 195–199.
- [10] 塔什甫拉提·尼扎木丁, 汪 昆, 艾斯卡尔·艾木都拉, 等. 统计与规则相结合的维吾尔语人名识别方法[J]. 自动化学报, 2017, 43(4): 653–664.
- [11] 热合木·马合木提, 于斯音·于苏普, 张家俊, 等. 基于模糊匹配与音字转换的维吾尔语人名识别[J]. 清华大学学报: 自然科学版, 2017, 57(2): 188–196.
- [12] 阿依古丽·哈力克, 艾山·吾买尔, 吐尔根·伊布拉音, 等. 汉维时间数字和量词的识别与翻译研究[J]. 中文信息学报, 2016, 30(6): 190–200.
- [13] 杨 攀, 李 森, 张 建. 基于短语统计翻译的汉维机器翻译系统[J]. 计算机应用, 2009, 29(7): 2022–2025.
- [14] 董兴华, 周俊林, 郭树盛, 等. 基于短语的汉维/维汉统计机器翻译[J]. 计算机工程, 2011, 37(9): 16–18.
- [15] 阿米妮古丽·奥斯曼, 加日拉·买买提热依木, 吐尔根·依布拉音. 维汉/汉维机器翻译后编辑器的设计与实现[J]. 新疆大学学报: 自然科学版, 2013, 30(4): 444–450.
- [16] 贾志先. 维吾尔语口语考试系统的开发与应用[J]. 计算机技术与发展, 2015, 25(5): 205–208.
- [17] BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263–311.
- [18] VOGEL S, NEY H, TILLMANN C. HMM-based word alignment in statistical translation[C]//Proceedings of the 16th conference on computational linguistics. Copenhagen, Denmark; Association for Computational Linguistics, 1996: 836–841.
- [19] OCH F J, NEY H. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2003, 29(1): 19–51.
- [20] LIU Yang, XIA Tian, XIAO Xinyan, et al. Weighted alignment matrices for statistical machine translation[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. Singapore; Association for Computational Linguistics, 2009: 1017–1026.
- [21] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]//Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology. Edmonton, Canada; Association for Computational Linguistics, 2003: 48–54.
- [22] JAEGER H. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach[R]. [s. l.]: [s. n.], 2002.
- [23] GUTHRIE D, ALLISON B, LIU Wei, et al. A closer look at skip-gram modelling[C]//Proceedings of the 5th international conference on language resources and evaluation. Genoa, Italy; European Language Resources Association, 2006: 1–4.
- [24] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. Lake Tahoe; Neural Information Processing Systems Foundation, 2013: 3111–3119.
- [25] MCCALLUM A, NIGAM K. A comparison of event models for naive bayes text classification[C]//Workshop on learning for text categorization. Madison, Wisconsin; Association for the Advancement of Artificial Intelligence, 1998: 41–48.
- [26] OCH F J. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st annual meeting on association for computational linguistics. Sapporo, Japan; Association for Computational Linguistics, 2003: 160–167.
- [27] PAPINENI K, ROUKOS S, WARD T, 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.
- +++++
- (上接第 143 页)
- [6] CHENG Penggen, ZHOU Guoqing, ZHENG Zezhong. Detecting and counting vehicles from small low-cost UAV images[C]//ASPRS. Baltimore: [s. n.], 2009: 9–13.
- [7] 张专成, 张孝杰, 邹 涛. 用于数字图像直方图处理的一种二值映射规则[J]. 中国图象图形学报, 2004, 9(3): 280–284.
- [8] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded up robust features[C]//Proceedings of the 9th European conference on computer vision. Graz, Austria; Springer-Verlag, 2006: 404–417.
- [9] CHUM O, MATAS J, KITTLER J. Locally optimized RANSAC[C]//Joint pattern recognition symposium. Berlin; Springer, 2003: 236–243.
- [10] 牛 彦. 关于透视变换的研究[J]. 计算机辅助设计与图形万方数据学报, 2001, 13(6): 549–551.
- [11] 江玉林, 陈学平, 李振宇, 等. 公路路域环境区域划分与环境特征的调查研究[C]//草坪与地被科学进展论文汇编. 北京: 中国林业出版社, 2006.
- [12] 陆宗骥, 梁 诚. 用 Sobel 算子细化边缘[J]. 中国图象图形学报, 2000, 5(6): 516–520.
- [13] 吴 捷, 陈德智, 郭成志. Sobel 边缘检测算法的变异实现图像增强[J]. 激光与红外, 2008, 38(6): 612–614.
- [14] 安 如, 冯学智, 王慧麟. 基于数学形态学的道路遥感影像特征提取及网络分析[J]. 中国图象图形学报, 2003, 8(7): 798–804.
- [15] 李朝锋, 潘婷婷. 基于形态学开闭运算和梯度优化的分水岭算法的目标检测方法[J]. 计算机应用研究, 2009, 26(4): 1593–1594.