翻译质量自动评价研究综述*

秦 颖
(北京外国语大学 计算机系, 北京 100089)

摘 要: 随着机器翻译研究的推进和翻译教学方式的革新,译文质量自动评价问题近年来受到大量关注。为把握翻译质量自动评价的思路、方法,通过对目前研究脉络的梳理,从研究特点角度绘制出了一个树型分类图谱,并对典型算法及其改进思路进行了分析;还对自动评价算法的评测方法、国际机器翻译评测平台和自动评测开放工具等给予了介绍。最后分析了当前研究存在的主要困难和问题,提出了对发展方向的展望。

关键词: 翻译质量; 自动评价; 质量估计; 算法
中图分类号: TP182 **文献标志码**: A **文章编号**: 1001-3695(2015)02-0326-04
doi:10.3969/j.issn.1001-3695.2015.02.002

Review on automatic translation quality evaluation

QIN Ying
(*Dept. of Computer Science, Beijing Foreign Studies University, Beijing* 100089, *China*)

Abstract: With the current development of machine translation and innovation of translation teaching, the issue of automatic translation quality evaluation has arisen a lot of concerns. In order to grasp the ideas and methods of translation evaluation, this paper proposed a systemic review on current researches. According to the characteristics of these studies, this paper drew a tree to illustrate the branches of different approaches. It also introduced typical algorithms and the map of their improvements, as well as the assessment on automatic evaluation, international shared task of machine translation evaluation and open toolkits of automatic evaluation. In the last section, it analyzed main obstacles and problems on current researches. It also put forward prospects on this field in the part.

Key words: translation quality; automatic evaluation; quality estimation; algorithm

0 引言

翻译研究必然伴随着翻译质量评价(translation quality evaluation or assessment),质量评价是翻译研究不可或缺的反馈环节。评价译文质量的应用需求十分广泛,不仅机器翻译系统需要评测和对比,在译文的出版编辑、语言翻译教学等领域也需要对译文的质量进行评价。目前评价翻译质量依然主要依赖人工,甚至是专家。译文质量评价是一个主观性比较强的问题,评分的高低实质是对评价者而言译文的可接受程度。同一个译文,不同的评价者及同一个评价者多次评价的结果并不完全一致(即 inter-and intra-agreement 问题)。

面对海量译文,人工评价越来越力不从心。自动评价因其快速、廉价、客观的特点吸引了众多的研究,尤其是在机器翻译研究蓬勃发展的今天,需要快速发现译文中的错误、调节翻译系统的参数、评价系统性能、进行不同系统的比较等,使得质量自动评价研究也成为热点。文献[1]绘制的机器翻译开发周期图(图 1)形象地描述了翻译评价的位置和作用。

2010 年,ACL 首次将翻译评价标准(metrics for machine translation)和机器翻译、系统综合一起列为统计机器翻译研讨(SMT workshop)的三大任务[2],为在共同的平台上研究自动评测方法提供便利。2010 年和 2013 年均有 14 支参赛队提交了几十种评测标准[2,3]。

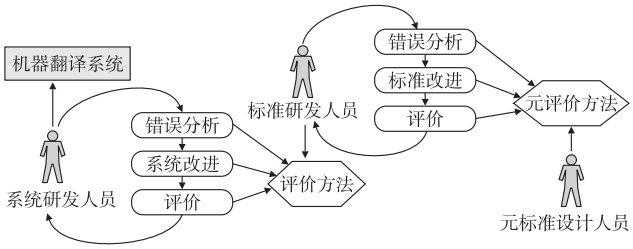

图 1 机器翻译开发周期图

语言教学和翻译出版领域的自动评价研究也有一些成果[4~6]。尽管待评译文不是出自机器而是人,但需求是类似的,都是对译文各方面的问题进行评价,即标志错误、评分等。

整体上,自动评价研究处于诸子百家的时代。尽管出现过几十种算法,也有开源的工具[1],但是和人工评价的相关度(correlation)都不高[2]。本文首先对目前的研究状况进行梳理,回顾了典型评价方法的发展,同时简单述及对评价算法的评测、国际自动评价研究平台和开放工具。总结了当前研究面临的困难和问题的讨论,以及未来的发展方向。

1 翻译评价研究分类

翻译自动评价的研究成果不断涌现,通过对目前掌握的文献进行梳理,本文从研究特点角度对其进行了分类。第一级分类是根据研究对象的不同进行划分,第二级是评价方式的不同,第三级是实现方法的不同,然后又从有无参考译文、评价粒

收稿日期: 2014-05-04; **修回日期**: 2014-06-20 **基金项目**: 校级资助项目(2009JJ056); 全国教育科学规划课题(GPA115033)
作者简介:秦颖(1971-),女,山东人,副教授,博士,主要研究方向为计算语言学(qinying@bfsu.edu.cn).

度、对语言知识的依赖程度等进一步细分。最后得到了一个研究分类的树型图谱(图 2),以期对目前的研究有个比较清晰的把握。

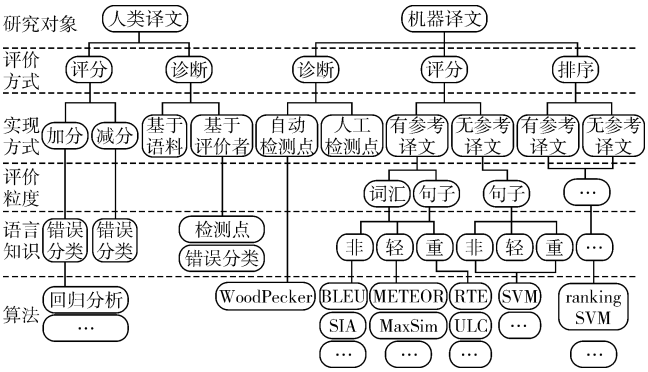

图 2　翻译自动评价研究分类图谱

首先根据译文的来源不同,分为人类译文(或学生译文)的评价和机器译文的评价两类,其中机器译文质量评价的研究成果居多。之所以存在这种分类状况,源于自动评价研究的文献普遍认为,评价人的译文要远远难于评价机器译文[7~9]。评测机器译文质量性能不错的算法用于评价人的译文时,不能区分人的译文中的细微差别[8]。文献[2]甚至指出,自动评测算法更适合的是统计机器翻译的评测,评价基于规则的机器翻译系统时会产生问题。因此,目前自动评价还分为两个研究领域。

1.1　人类译文的评价方法

人类译文的自动评价更多地模拟专家评价的思想。人类译文自动评价的评价方式包括评分和诊断两类,实现评价时通常有加分法和减分法两种。加分法是通过累计正确的得分点的分数来对译文打分,减分法则是基于译文中的错误从满分值中做减法。大多质量评价基于错误分类方案进行[4],即根据错误数及错误的严重程度实行减分。而错误分为两类:a)大错(major error),主要指译文基本成分的错,错误将导致语义混淆;b)小错(minor error),是使用了不正确或不恰当的表达方法或语法。当然,评价译文质量高低通常有一个人们对错误的容忍度问题,文献[10]认为 400 个词左右的译文最多允许有 12 个小错、1 个大错。因此关于错误体系的构建成为核心研究问题。美国翻译家协会 ATA 将错误划分为 22 种类型,不同类型的错有不同的分值;加拿大翻译局的 Sical 系统能识别 675 种错,包括 300 种词汇错和 375 种句法错;英国翻译与口译研究所 ITI 只有 18 类错误分类,而且每种错误分值相等。

对学习者译文评价的研究也在开展[8,9],研究以加分法为主。首先需要有专家参与确定译文的评分点,通过统计评分点的出现情况,并综合其他特征,如译文的形式特征、译文和原文的对齐特征等进行回归分析,得到译文的评分。

为了提供更客观的翻译反馈信息,文献[11]基于自然文本构建基准语料(benchmark),对学习者译文中的问题不仅评分,还提供更客观的翻译建议,比如以 KWIC(key word in context)的形式展示在实际语料中某种语言现象的使用情况等。研究的重点是提供真实的语言使用状况。

1.2　机器译文的评价方法

机器译文自动评价的研究近年来如雨后春笋,大致出现了诊断性评价(diagnostic evaluation)、评分(scoring)和排序(ranking)三种评价方式。

1.2.1　诊断性评价

诊断评价在上述三种方式中开展得最少,主要有文献[12,13]。文献[12]先由人工将测试句中重要语言测试点挑选出来并分类,然后在机器译文中自动检测这些测试点是否被正确译出,从而评价译文质量。测试点分为词语、成语、词法、基本语法、中级语法和高级语法等六类,分别设定对质量影响的权重,利用加分法进行评分。而文献[13]提出的用于“863”机器翻译评测的 WoodPecker,对检测点实现了自动提取,减少了对人工的依赖。

1.2.2　评分

评分是最多的自动评价方式。评价机器译文时,根据有无参考译文又分为两种研究。有参考译文的评价是通过将待评译文和参考译文作比较,根据相似程度评分,这种研究居多;不需要参考译文的评分也被称为质量估计(quality estimation)[14~17],根据译文特征将译文质量简单分为“好”或“坏”,或者区分人类译文(human-like)和非人译文(non-human-like)。质量估计被视做二分类问题,因此,支持向量机(SVM)等算法被用于了该种评价。还有一些研究介于有参考译文和无参考译文之间,比如文献[18]在没有人工参考译文的情况下,将若干机器译文生成伪参考译文(pseudo reference),然后用有参考译文的方法进行评测。

依赖参考译文的评价,参考译文就是标准答案,与参考译文越相似,译文质量越高,这个假设是评价算法的基本思想。而求待评译文和参考译文相似度的方法多种多样,这些方法根据语言粒度可以分为词汇层面的相似和句子/语篇层面的相似;根据对语言知识的依赖度也可分为非语言、轻语言和重语言[19,20]。

非语言的方法通常不需要语言层面的分析来计算相似,常见的有四种:基于编辑距离的方法,具体算法如 WER[21]、PER[22]、TER[23,24] 等;基于准确率的方法,如 BLEU[25]、NIST[7]、SIA[26]等;基于召回率的方法,如 ROUGE[27] 等;基于综合指标的方法,如 GTM[28]、PORT[29]等。

轻语言的方法需要利用一些语言信息进行质量评价,如词性 POS、同义词典等。著名的算法有 METEOR[30]、METEOR-NEXT[31]、TER-Plus[24]、MAXSIM[32]、wpBLE[33]、TESLA[19]、AMBER[20]等。

重语言的相似求解方法则对译文进行较多的语法或语义层面的分析,从句法结构(syntactic structure)、重述(paraphrase)、近义(synonym)、文本蕴涵(textual entailment)等语言方面计算待评译文和参考译文的相似度,如 ULC[1]、RTE[34]、DCU-LFG[35] 等都需要对译文作较深入的分析和处理,评价的代价高。

1.2.3　排序评价法

排序法(ranking)[8,36] 适用于对一组译文进行评价,根据质量高低排序。文献[37]认为排序评价的优势有三点:a)人工评测时,排序比打分更容易;b)人工排序评价的一致性比打分一致性更高;c)更适用于系统之间的比较。文献[8]将 BLEU 得分、依存关系匹配、困惑度(perplexity)融合到 ranking SVM 学习方法中,根据 SVM 的得分对一组机器译文的优劣排序。

总之,上述各种算法都试图对译文的质量进行区分,但影响译文质量的因素是多方面的,常见的包括译文的流利度(flu-

ency)和充分性(adequacy),有时还有其他因素,如可理解性(understandability)。上述方法往往是多个因素综合评价的结果。如果分开评测时会发现,这些算法对充分性的评价性能更好些,流利度指标更难评[38],这也是目前机器译文和人类译文的最大区别。所以有的学者专注于句子流利度的评价方法,因为人的译文都比机器译文通顺得多。文献[9,39]发现,句法结构信息更有利于抓住流利的本质;文献[40]则研究了与流利有关的错误类别划分。

2 典型评测算法的思想和发展

2.1 BLEU

翻译自动评价中,BLEU[25]是必然要提及的算法。BLEU的影响是划时代的,尽管在此之前也有类似的观点。当前BLEU仍作为评价的一个基准(benchmark)。BLEU的基本假设是,如果待评译文和参考译文共现的N-gram越多,说明越相似,译文质量就越高。通过统计共现的N-gram数目,并对短句增加惩罚因子,就可以借助参考译文对同一个题目的译文进行评测。BLEU得分的计算公式很简单。

$$\text{score}_{\text{BLEU}} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

其中:BP是对长度小于参考值的译文的惩罚因子,p_n是N-gram匹配率。

机械匹配策略和N-gram的稀疏问题(尤其是大于bigram的数目十分稀疏)是BLEU的主要缺点[38,41]。文献[42]甚至指出BLEU算法其实不足以反映译文质量的改变。针对上述缺点很多学者提出了改进方法,如M-BLEU[43]、EBLEU[44]、AMBER[20]等。其中AMBER中设置了10种惩罚因子,每一种惩罚因子人工赋予不同的权重;并且采用包括公共最长字串LCS等四种与参考译文匹配的策略;有八种类型的文本作为输入,可谓是对BLEU算法的极致改进。

2.2 METEOR

以METEOR为代表的算法的主要思想是基于词汇的相似度进行评价,利用WordNet等少量外部资源增加同义词的匹配几率。先后出现过METEOR-NEXT[31]、TESLA[19]等。TESLA的改进之处是对N-gram加权,加入了词序相似等。还有一些算法是融合METEOR和其他算法得到的,如SIA[26]结合了METEOR和ROUGE-W[27],在柔性匹配基础上,更考虑了对齐词之间的距离因素。

纵观基于N-gram匹配评价译文的算法发展思路,基本是从两个角度对其改进,一是采用更柔性的匹配策略或利用外部语言资源,扩大译文和参考译文相似的检测范围;二是多种相似函数的利用,相似函数涉及字面相似、距离相似、语义相似等,然后对各种相似值进行综合,或线性组合,或求各种均值(算术平均、几何平均、调和平均)[19]。

3 自动评价算法的评测及其他

3.1 自动评价算法的评测

有了各种评价算法,自然就需要对算法进行比较。最常见的方法是将自动评价和人工评价作相关分析,常用的有Pearson相关系数、Spearman相关系数和Kendall[45]。另外还有ORANGE分值[46]。目前还没有机器译文质量优于参考译文,所以如果将参考译文和机器译文混在一起由算法评价时,评价方法至少应该将参考译文的质量排在机器译文前面,因此,参考译文的排名率(ranking ratio)越小,也就能说明评价算法性能越好,这就是ORANGE分值的含义。

对诊断性评价,目前还缺乏合适的评测方法。

3.2 国际评测平台和自动评测开放工具

自WMT 10开始,增设了评价标准的研究平台。在WMT 13的质量估计任务(quality estimation)中包括两个任务,一个是句子级的评测,另一个是词语级的评测。研究数据全部公开,方便比较。

一般在机器翻译研究项目中同时公开评测工具。国际机器翻译研究平台有美国标准和技术研究院NIST(www.nist.gov)、欧盟的TC-STAR等,国内中科院也组织国内的机器翻译研究和评测。最近推出的开放评测工具如Asiya[1],提供了进入各种评测方法的公共界面(http://www.lsi.upc.edu/~nlp/Asiya)。QUEST[17]是一个质量估计的框架,有各种提取原文、译文和外部资源特征的工具(http://www.quest.dcs.shef.ac.uk/)。

4 研究存在的主要困难和趋势展望

从以上对文献的分类和回顾不难看出:翻译质量评价主要基于语言的浅层面进行,与参考译文基于n-gram的匹配是基本出发点,所以评测算法不能完全反映译文质量的改变就不难理解了。参考译文只能提供部分正确的译法,对其他合理的译文,算法无法进行泛化判断。自动评价算法和人工评价的相关度普遍较低,尚不能替代人工评测,自动评价研究亟待深入。

对于自动评价下一步的研究方向,也是热议的问题。文献[36]认为基于译文内容的评价是趋势;文献[47]开始关注没有参考译文的评价;更有一些学者试图从语义层面上分析译文间的相似。当然,评价方法的稳定性、译文质量改变时的敏感性等也是要考虑的问题[34]。

笔者认为,翻译质量自动评价是一个多层体系,从质量排序到评分,再到诊断评价,评价粒度越来越细。在这个框架体系下,评价应该具有一个统一的平台,不仅能够评价学生译文,也能评价机器译文。评价时,参考译文是一种参照,另外,原文的信息也应成为评价的重要依据。除了语言形式上的相似比较,更多地从语言信息角度关注译文的内容和结构,构建系统的语言知识库,才能从较深层次上进行质量评价。随着大规模语料库的建成,自然语言大数据也将对自动评价方法产生影响。

翻译自动评价的应用需求广泛,从机器翻译研究到翻译教学、出版编辑,都希望廉价、快速、准确、客观的自动评价方法能够替代或部分替代人工评价。翻译自动评价研究虽然起步较晚,但新的思想和探索不断涌现,文献丰富。由于笔者掌握的文献有限,观点难免有失偏颇,希望通过本文的梳理,对翻译自动评价的研究提供参考和帮助。

致谢 感谢北京外国语大学中国外语教育研究中心对本研究的资助。

参考文献:

[1] GIMÉNEZ J, MÀRQUEZ L. Asiya: an open toolkit for automatic machine translation (meta-) evaluation[J]. The Prague Bulletin of Mathematical Linguistics, 2010,94:77-86.

[2] CALLISON-BURCH C, KOEHN P, MONZ C, *et al*. Findings of the

2010 joint workshop on statistical machine translation and metrics for machine translation[C]//Proc of the 5th Joint Workshop on Statistical Machine Translation and MetricsMATR. Stroudsburg: Association for Computational Linguistics, 2010: 17-53.

[3] BOJAR O, BUCK C, CALLISON-BURCH C, *et al*. Findings of the 2013 workshop on statistical machine translation[C]//Proc of the 8th Workshop on Statistical Machine Translation. 2013:1-44.

[4] SECARA A. Translation evaluation: a state of the art survey[C]// Proc of eCoLoRe/MeLLANGE Workshop. 2005: 39-44.

[5] 王金铨. 中国学习者汉译英机助评分模型的构建[D]. 北京:北京外国语大学, 2008.

[6] 秦颖,文秋芳. 大规模考试英汉互译自动评分系统的研发与应用[M]. 北京:高等教育出版社, 2012.

[7] DODDINGTON G. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics[C]// Proc of the 2nd International Conference on Human Language Technology Research. San Francisco: Morgan Kaufmann Publishers, 2002: 138-145.

[8] YE Yang, ZHOU Ming, LIN C Y. Sentence level machine translation evaluation as a ranking problem: one step aside from BLEU[C]// Proc of the 2nd Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2007: 240-247.

[9] CHAE J, NENKOVA A. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text[C]// Proc of the 12th Conference of the European Chapter of the ACL. 2009:139-147.

[10] MALCOM W. The application of argumentation theory to translation quality assessment[J]. Meta, 2001,46(2):327-344.

[11] BOWKER L. Towards a methodology for a corpus-based approach to translation evaluation[J]. Meta, 2001, 46(2):345-364.

[12] YU Shi-wen. Automatic evaluation of output quality for machine translation systems[J]. Machine Translation, 1993,8(1-2): 117-126.

[13] ZHOU Ming, WANG Bo, LIU Shu-jie, *et al*. Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points[C]//Proc of the 22nd International Conference on Computational Linguistics. 2008:1121-1128.

[14] BLATZ J, FITZGERALD E, FOSTER G, *et al*. Confidence estimation for machine translation[R]. Baltimore: Johns Hopkins University, 2003.

[15] GANDRABUR S, FOSTER G. Confidence estimation for translation prediction[C]//Proc of the 7th Conference on Natural Language Learning. 2003:95-102.

[16] GAMON M, AUE A, SMETS M. Sentence-level MT evaluation without reference translations: beyond language modeling[C]//Proc of the 10th European Association for Machine Translation Conference. 2005.

[17] SPECIA L, RAJ D, TURCHI M. Machine translation evaluation versus quality estimation[J]. Machine Translation, 2010, 24(1): 39-50.

[18] ALBRECHT J, HWA R. Regression for sentence-level MT evaluation with pseudo references[C]//Proc of the 45th Meeting of the Association for Computational Linguistics. 2007: 296-303.

[19] LIU Chang, DAHLMEIER D, NG H T. TESLA: translation evaluation of sentences with linear-programming-based analysis[C]// Proc of the 5th Joint Workshop on Statistical Machine Translation and MetricsMATR. Stroudsburg: Association for Computational Linguistics, 2010: 354-359.

[20] CHEN Bo-xing, KUHN R. AMBER: a modified BLEU, enhanced ranking metric[C]//Proc of the 6th Workshop on Statistical Machine Translation. 2011: 71-77.

[21] NIEβEN S, OCH F J, LEUSCH G, *et al*. An evaluation tool for machine translation: fast evaluation for MT research[C]//Proc of the 2nd International Conference on Language Resources and Evaluation. 2000.

[22] LEUSCH G, UEFFING N, NEY H. A novel string-to-string distance measure with applications to machine translation evaluation[C]// Proc of MT Summit IX. 2003.

[23] SNOVER M, DORR B, SCHWARTZ R, *et al*. A study of translation edit rate with targeted human annotation[C]//Proc of Conference on the Association for Machine Translation in Americas. 2006.

[24] SNOVER M, MADNANI N, DORR B, *et al*. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric[C]// Proc of WMT Workshop. 2009.

[25] PAPINENI K, ROUKOS S, WARD T, *et al*. BLEU: a method for automatic evaluation of machine translation[C]// Proc of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 311-318.

[26] LIU Ding, GILDEA D. Stochastic iterative alignment for machine translation evaluation[C]// Proc of the 21st COLING/ACL Conference on Computational Linguistics. 2006:39-546.

[27] LIN C Y, OCH F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics[C]// Proc of the 42nd Annual Meeting of the Association for Computational Linguistics. 2004.

[28] TURIAN J P, SHEN L, MELAMED I D. Evaluation of machine translation and its evaluation[C]//Proc of MT Summit IX. 2003.

[29] CHEN Bo-xing, KUHN R, LARKIN S. PORT: a precision-order-recall MT evaluation metric for tuning[C]// Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 930-939.

[30] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]// Proc of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005: 65-72.

[31] DENKOWSKI M, LAVIE A. Extending the METEOR machine translation evaluation metric to the phrase level[C]//Proc of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 250-253.

[32] CHAN Y S, NG H T. MAXSIM: a maximum similarity metric for machine translation evaluation[C]//Proc of ACL Meeting. 2008: 55-62.

[33] POPOVI′C M, NEY H. Syntax-oriented evaluation measures for machine translation output[C]// Proc of WMT Workshop. 2009.

[34] PADÓ S, GALLEY M, JURAFSKY D, *et al*. Robust machine translation evaluation with entailment features[C]// Proc of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2009: 297-305.

[35] HE Yi-fan, DU Jin-hua, WAY A, *et al*. The DCU dependency-based metric in WMTMetricsMATR 2010[C]//Proc of the 5th Joint Workshop on Statistical Machine Translation and Metrics MATR. 2010: 324-328.

[36] AVRAMIDIS E. Comparative quality estimation: automatic sentence-level ranking of multiple machine translation outputs[C]//Proc of the 24th International Conference on Computational Linguistics. 2012: 115-132.

[37] DUH K. Ranking vs. regression in machine translation evaluation[C]// Proc of the 3rd Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2008: 191-194.

(下转第 335 页)

[23] RADMEHR A, GHASEMI A. A novel video error concealment technique using modified boundary matching algorithm with correlation function[C]//Proc of the 21st Iranian Conference on Electrical Engineering. 2013: 1-4.

[24] CHEN L Y, CHAN S C, SHUM H Y. A joint motion-image inpainting method for error concealment in video coding[C]//Proc of IEEE International Conference on Image Processing. 2006: 2241-2244.

[25] WANG Yi, GUO Xiao-qiang, FENG Ye, *et al*. A novel temporal error concealment framework in H. 264/AVC[C]//Proc of International Conference on IEEE Multimedia and Expo. 2013: 1-6.

[26] ZHAO Bin, DELP E J. Inter-layer error concealment for scalable video coding based on motion vector averaging and slices interleaving [C]//Proc of International Conference on IEEE Multimedia and Expo. 2013: 1-6.

[27] ZHAO Chen, MA Si-wei, ZHANG Jian, *et al*. A highly effective error concealment method for whole frame loss[C]//Proc of IEEE International Symposium on Circuits and Systems. 2013: 2135-2138.

[28] WANG P C, LIN C S. Enhanced backward error concealment for H. 264/AVC videos on error-prone networks[C]//Proc of International Conference on Biometrics and Security Technologies. 2013: 62-66.

[29] LEE Wen-nung, LEE Chang-ming, GAO Zhi-wei, *et al*. Motion vector recovery for video error concealment by using iterative dynamic-programming optimization[J]. IEEE Trans on Multimedia, 2013, 16(1): 216-227.

[30] SEILER J, SCHOBERL M, KAUP A. Spatio-temporal error concealment in video by de-noised temporal extrapolation refinement [C]// Proc of International Conference on IEEE Image Processing. 2013: 1613-1616.

[31] CHEN Xiao-ming, CHUNG Y Y, BAE C. Dynamic multi-mode switching error concealment algorithm for H. 264/AVC video applications[J]. IEEE Trans on Consumer Electronics, 2008, 54(1): 154-162.

[32] ZHANG Da-qing, LI Sheng-hong, YANG Kong-jin, *et al*. An error concealment adaptive framework for intra-frames [C]//Proc of International Conference on IEEE Image Processing. 2013: 1880-1884.

[33] WANG Yi, GUO Xiao-qiang, FENG Ye, *et al*. A novel temporal error concealment framework for H. 264 over wireless networks[C]// Proc of International Conference on IEEE Wireless Personal Multimedia Communications. 2013: 1-5.

[34] LIN Qi-wei, MAO Yuan. Adaptive MB size selecting based video error concealment algorithm[C]//Proc of International Conference on Anti Counterfeiting, Security and Identification. 2012: 1-4.

[35] SULLIVAN G J, OHM J, HAN W J, *et al*. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Trans on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

[36] NIGHTINGALE J, WANG Qi, GRECOS C. HEVC stream: a framework for streaming and evaluation of high efficiency video coding (HEVC) content in loss-prone networks [J]. IEEE Trans on Consumer Electronics, 2012, 58(2): 404-412.

[37] LIN Ting-lan, YANG N C, SYU R H, *et al*. Error concealment algorithm for HEVC coded video using block partition decisions[C]// Proc of International Conference on Communication and Computing, Signal Processing. 2013: 1-5.

[38] CHANG Y L, REZNIK Y A, CHENG Zhi-feng, *et al*. Motion compensated error concealment for HEVC based on block-merging and residual energy[C]//Proc of the 20th International Packet Video Workshop. 2013: 1-6.

[39] 刘畅,马然,刘德阳,等. HEVC 中基于前后景区域的错误隐藏[J]. 电视技术, 2012, 36(15): 8-11.

[40] 金惠美. 浅谈下一代编码压缩技术——HEVC[J]. 数字通信世界, 2012(11): 62-64.

[41] CHUNG T Y, SULL S, KIM C S. Frame loss concealment for stereoscopic video plus depth sequences [J]. IEEE Trans on Consumer Electronics, 2011,57(3):1336 - 1344.

[42] LIN Ting-lan, CHANG T E, HUANG G S, *et al*. Multi-view video error concealment with improved pixel estimation and illumination compensation[C]//Proc of International Conference on Intelligent Signal Processing and Communications Systems. 2013: 157-162.

[43] 时琳,刘荣科,李君烨. 基于深度信息的立体视频错误隐藏方法[J]. 电子与信息学报, 2012, 34(7): 1678-1684.

[44] LIU Shu-jie, CHEN Ying, WANG Ye-kui, *et al*. Frame loss error concealment for multi-view video coding[C]//Proc of International Symposium on IEEE Circuits and System. 2008: 3470-3473.

[45] 周洋,蒋刚毅,郁梅,等. 面向 HBP 编码格式的立体视频 B 帧整帧丢失分层错误隐藏算法[J]. 电子与信息学报, 2014, 36(2): 377-383.

(上接第 329 页)

[38] LIU Ding, GILDEA D. Syntactic features for evaluation of machine translation[C]// Proc of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization. 2005:25-32.

[39] MUTTON A, DRAS M, WAN S, *et al*. GLEU: automatic evaluation of sentence-level fluency[C]//Proc of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:344-351.

[40] ELLIOTT D, HARTLEY A, ATWELL E. Fluency error categorization scheme to guide automated machine translation evaluation[M]//Machine Translation: from Real Users to Research. Berlin: Springer-Verlag, 2004: 64-73.

[41] CULY C, RIEHEMANN S Z. The limits of N-gram translation evaluation metrics[C]//Proc of MT Summit IX. 2003: 71-78.

[42] CALLISON-BURCH C, KOEHN P, OSBORNE M. Improved statistical machine translation using paraphrases[C]//Proc of Human Language Technology Conference of the North American Chapter of the ACL. 2006.

[43] AGARWAL A, LAVIE A. METEOR, M-BLEU and M-TER: evaluation metrics for high-correlation with human rankings of machine translation output[C]//Proc of the 3rd Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2008: 115-118.

[44] HAN A L F, LU Yi, WONG D F, *et al*. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling[C]// Proc of the 8th ACL Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2013: 365-372.

[45] ENRIQUE A, GONZALO J, PENAS A, *et al*. QARLA: a framework for the evaluation of automatic summarization[C]//Proc of the 43rd Annual Meeting of the Association for Computational Linguistics. 2005:280-289.

[46] LIN C Y, OCH F J. ORANGE: a method for evaluating automatic evaluation metrics for machine translation[C]//Proc of the 20th International Conference on Computational Linguistics. 2004:501-507.

[47] SPECIA L, SHAH K, De SOUZA J G C, *et al*. QuEst: a translation quality estimation framework[C]//Proc of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 79-84.