软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(4):1051−1066 [doi: 10.13328/j.cnki.jos.006221]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向神经机器翻译系统的多粒度蜕变测试*

钟文康[1], 葛季栋[1], 陈 翔[2], 李传艺[1], 唐 泽[1], 骆 斌[1]

[1](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

[2](南通大学 信息科学技术学院,江苏 南通 226019)

通讯作者: 李传艺, E-mail: lcy@nju.edu.cn

摘 要: 机器翻译是利用计算机将一种自然语言转换成另一种自然语言的任务,是人工智能领域研究的热点问题之一.近年来,随着深度学习的发展,基于序列到序列结构的神经机器翻译模型在多种语言对的翻译任务上都取得了超过统计机器翻译模型的效果,并被广泛应用于商用翻译系统中.虽然商用翻译系统的实际应用效果直观表明了神经机器翻译模型性能有很大的提升,但如何系统地评估其翻译质量仍是一项具有挑战性的工作.一方面,若基于参考译文评估翻译效果,其高质量参考译文的获取成本非常高;另一方面,与统计机器翻译模型相比,神经机器翻译模型存在更显著的鲁棒性问题,然而还没有探讨神经机器翻译模型鲁棒性的相关研究.面对上述挑战,提出了一种基于蜕变测试的多粒度测试框架,用于在没有参考译文的情况下评估神经机器翻译系统的翻译质量及其翻译鲁棒性.该测试框架首先在句子粒度、短语粒度和单词粒度上分别对源语句进行替换,然后将源语句和替换后语句的翻译结果进行基于编辑距离和成分结构分析树的相似度计算,最后根据相似度判断翻译结果是否满足蜕变关系.分别在教育、微博、新闻、口语和字幕这 5 个领域的中英文数据集上对 6 个主流商用神经机器翻译系统使用不同的蜕变测试框架进行了对比实验.实验结果表明,所提方法在与基于参考译文方法的皮尔逊相关系数和斯皮尔曼相关系数上分别比同类型方法高 80%和 20%,说明提出的无参考译文的测试评估方法与基于参考译文的评估方法的正相关性更高,验证了其在评估准确性上显著优于同类型其他方法.

关键词: 神经网络;机器翻译;质量评估;蜕变测试;多粒度

中图法分类号: TP311

中文引用格式: 钟文康,葛季栋,陈翔,李传艺,唐泽,骆斌.面向神经机器翻译系统的多粒度蜕变测试.软件学报,2021, 32(4):1051−1066. http://www.jos.org.cn/1000-9825/6221.htm

英文引用格式: Zhong WK, Ge JD, Chen X, Li CY, Tang Z, Luo B. Multi-granularity metamorphic testing for neural machine translation system. Ruan Jian Xue Bao/Journal of Software, 2021,32(4):1051−1066 (in Chinese). http://www.jos.org.cn/1000-9825/6221.htm

Multi-granularity Metamorphic Testing for Neural Machine Translation System

ZHONG Wen-Kang[1], GE Ji-Dong[1], CHEN Xiang[2], LI Chuan-Yi[1], TANG Ze[1], LUO Bin[1]

[1](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

[2](School of Information Science and Technology, Nantong University, Nantong 226019, China)

Abstract: Machine translation task focuses on converting one natural language into another. In recent years, neural machine translation models based on sequence-to-sequence models have achieved better performance than traditional statistical machine translation models on multiple language pairs, and have been used by many translation service providers. Although the practical application of commercial translation system shows that the neural machine translation model has great improvement, how to systematically evaluate its translation

* 基金项目: 国家自然科学基金(61802167, 61972197, 61802095); 江苏省自然科学基金(BK20201250)

Foundation item: National Natural Science Foundation of China (61802167, 61972197, 61802095); Natural Science Foundation of Jiangsu Province of China (BK20201250)

本文由“面向领域的软件系统构造与质量保障”专题特约编辑潘敏学教授、魏峻研究员、崔展齐教授推荐.

收稿时间: 2020-09-12; 修改时间: 2020-10-26; 采用时间: 2020-12-19; jos 在线出版时间: 2021-01-22

quality is still a challenging task. On the one hand, if the translation effect is evaluated based on the reference text, the acquisition cost of high-quality reference text is very high. On the other hand, compared with the statistical machine translation model, the neural machine translation model has more significant robustness problems. However, there are no relevant studies on the robustness of the neural machine translation model. This study proposes a multi-granularity test framework MGMT based on metamorphic testing, which can evaluate the robustness of neural machine translation systems without reference translations. The testing framework first replaces the source sentence on sentence-granularity, phrase-granularity, and word-granularity respectively, then compares the translation results of the source sentence and the replaced sentences based on the constituency parse tree, and finally judges whether the result satisfies the metamorphic relationship. The experiments are conducted on multi-field Chinese-English translation datasets and six industrial neural machine translation systems are evaluated, and compared with same type of metamorphic testing and methods based on reference translations. The experimental results show that the proposed method MGMT is 80% and 20% higher than similar methods in terms of Pearson's correlation coefficient and Spearman's correlation coefficient respectively. This indicates that the non-reference translation evaluation method proposed in this study has a higher positive correlation with the reference translation based evaluation method, which verifies that MGMT's evaluation accuracy is significantly better than other methods of the same type.

Key words: neural network; machine translation; quality estimation; metamorphic test; multi-granularity

1 引 言

机器翻译研究如何将基于一种自然语言描述的文本自动翻译成基于另一种自然语言描述的文本,是自然语言处理的一个重要研究问题.传统的机器翻译系统主要采用统计机器翻译模型[1].近年来,随着深度学习的发展和应用,基于序列对序列模型(sequence to sequence model)的神经机器翻译模型[2]在很多语言对的机器翻译任务上都超过了统计机器翻译模型.神经机器翻译模型不仅有很高的研究价值,还有很强的产业化能力[3],目前主流的翻译服务提供商(例如,谷歌翻译、必应翻译、百度翻译、腾讯翻译等)都提供了在线神经机器翻译服务.

尽管神经机器翻译为机器翻译任务带来了极大的性能上的提升,但仍存在一些问题.例如,对长句子和低频词语的翻译效果不佳,翻译结果和词对齐模型不符等[4],并且,这些错误出现的规律和原因往往难以被发现.与统计机器翻译模型相比,神经机器翻译系统还存在更为显著的鲁棒性问题[4].Cheng 等人[5]指出,对输入语句做出的极小改变可能会引起翻译结果的剧烈改变,如同"蝴蝶效应".此外,目前商用的神经机器翻译系统较多,但由于神经网络模型结构和训练数据的差异,各神经机器翻译系统的稳定性并不一样.图 1 和图 2 分别展示了谷歌和百度的神经机器翻译系统在翻译 3 个近似句子时的不同结果.3 个待翻译英文句子在结构上完全相同,在内容上仅句尾单词含义不同,但是它们经过谷歌和百度的神经机器翻译系统翻译得到的翻译结果却出现了较大的差异.谷歌翻译在第 2 个和第 3 个句子上出现了翻译错误,而百度翻译在第 1 个句子上出现了翻译错误.

Fig.1 Translation errors of Google's neural machine translation system

图 1 谷歌神经机器翻译系统的翻译错误示例

Fig.2 Translation errors of Baidu's neural machine translation system

图 2 百度神经机器翻译系统的翻译错误示例

© 中国科学院软件研究所 http://www.jos.org.cn

显然,对神经机器翻译系统进行翻译鲁棒性评估具有重要的现实意义和研究意义.目前在该领域还缺乏相关研究.传统的机器翻译质量评估通常采用基于参考译文的方法,侧重翻译的正确性.而神经机器翻译系统采用的神经网络模型导致其与传统的统计机器翻译模型相比存在更为显著的翻译鲁棒性问题,亟需合理的测试手段和评估指标.如何对神经机器翻译系统进行测试和质量评估具有一定的研究挑战性.目前,这种挑战性主要体现在两个方面.

神经网络模型的测试困难性.神经机器翻译系统采用的神经网络模型缺乏可解释性和可理解性[6].在进行神经机器翻译时,待翻译语句在神经网络结构中会被转换为多维向量,这种转换涉及到的步骤繁杂,参数众多,很难理解每一个步骤的实际含义.另一方面,神经网络模型对训练数据具有很强的依赖性,相同的网络结构在不同的训练数据集下,训练出的参数取值会存在较大差异,造成输出的稳定性较低.

机器翻译任务的评估困难性.翻译质量通常基于参考译文进行评估,即给定人工翻译结果,与神经机器翻译系统输出的翻译进行比较,并通过相似度指标进行量化.但这种方法完全依赖于参考译文的质量,而高质量的参考译文获取的难度较大,成本很高.

神经网络模型的测试困难性意味着采用白盒测试方法可行性较低,而机器翻译模型的黑盒测试方法通常基于参考译文,成本较高.为了解决上述研究挑战,实现在没有参考译文情况下对神经机器翻译系统进行有效的翻译鲁棒性评估,本文基于蜕变测试思想提出了一个多粒度的蜕变测试框架 MGMT(multi-granularity metamorphic test).MGMT 首次采用多粒度的蜕变测试方法进行质量评估,分别在句子、短语和单词粒度上定义了蜕变关系以及相似度计算方法,并基于蜕变关系对每一个句子进行 3 个粒度上的蜕变测试,最后用蜕变关系满足率作为神经机器翻译系统的鲁棒性量化指标.同时,我们基于 MGMT 框架开展了实证研究,采用一个公开的中英翻译数据集 UM-Corpus[7],选取其中 5 个领域(教育、微博、新闻、口语、字幕)的英文句子集作为源数据集,在 MGMT 测试框架下对现有的、使用广泛的大型神经机器翻译系统(包括谷歌翻译[8]、必应翻译[9]、百度翻译[10]、阿里巴巴翻译[11]、腾讯翻译[12]、搜狗翻译[13])进行质量评估.最后将数据集中的中文句子作为参考译文,以基于参考译文的方法为基准,与同类型的蜕变测试方法进行比较,以证明 MGMT 相比于同类型方法在评估准确度上有显著的优越性.

本文第 2 节对已有的面向神经机器翻译系统的质量评估和测试工作进行总结.第 3 节介绍本文提出的多粒度蜕变测试框架,描述测试流程、蜕变关系定义以及相似度计算方法.第 4 节针对 6 个主流商用神经机器翻译系统在一个多领域的翻译数据集上进行实验,并用同类型蜕变测试工作与基于参考译文的测试方法进行对比,以证明本文方法的有效性.第 5 节进行总结与展望,对本文工作进行总结并阐明未来可能的工作方向.

2 相关工作

传统的机器翻译系统质量评估并不区分正确性和鲁棒性,通常用翻译质量来衡量系统质量.Eirini[14]总结了两类翻译质量评估方法.一类是人工评估,即由专业译者来判断翻译质量的好坏.人工评估的优点是评估结果最接近实际,但是时间成本和人力成本都较高.另一类方法是基于参考译文进行评估,即给定翻译好的参考译文,将机器翻译的输出结果与参考译文进行相似度指标计算,最常用的指标有 BLEU[15]、METEOR[16]、WER[17]等.基于参考译文的方法相对于人工方法成本有所降低,但是高质量参考译文的获取难度较大,成本仍然很高.

如何在没有参考译文的情况下对神经机器翻译系统进行质量评估是一项困难的任务.神经机器翻译系统采用的神经网络模型具有参数规模大、可理解性弱的特点,且普遍存在测试预言问题.测试预言问题[18]是指在测试中对于某个输入需要给定预期的输出来判断系统实际输出的正确性.Wang 等人[19]总结了目前常见的解决深度神经网络系统测试预言的方法,将其分为两类.第 1 类基于差异测试[20],即通过检测同一输入在基于相同规约的实现下的输出是否相同来判断是否出错.另一类基于蜕变测试[21],即通过定义蜕变关系来描述系统的输入变化和输出变化之间的关系.在以往的神经机器翻译系统质量评估工作中,基于蜕变测试的方法较为常见,这种方法的关键在于蜕变关系的定义.

Milam 等人[22]提出用往返翻译 RTT(round-trip translation)在无需参考译文的情况下可以对机器翻译系统

© 中国科学院软件研究所 http://www.jos.org.cn

进行测试的有效性.基于 RTT 构造的蜕变关系是:源语句通过神经机器翻译系统翻译到目标语言,再翻译回源语言得到的翻译结果应该与源语句相同.

Daniel 等人[23]提出了一种结合蒙特卡洛随机算法和蜕变测试的方法 MCMT(Monte Carlo combined metamorphic test),以此来衡量神经机器翻译系统的质量.它定义了一种类似 RTT[22]的蜕变关系:源语言经过神经机器翻译系统直接翻译到目标语言,与源语言先使用蒙特卡洛算法随机翻译到一种中间语言,再翻译到目标语言得到的两个翻译结果应该相同.

Zhou 等人[24]在 Daniel 等人[23]工作的基础上提出了新的神经机器翻译系统质量评估方法 MT4MT.该方法使用基于词替换的蜕变关系:替换源语句中的一个单词,不会影响翻译语句的结构.同时,MT4MT 有针对性地设计了一些简单的替换规则.

此外,有部分工作研究如何在无需参考译文的情况下进行机器翻译系统的翻译错误定位.He 等人[25]提出了结构不变性测试(structure-invariant test,简称 SIT)以发现系统的翻译错误.结构不变的含义是指,上下文含义相近的句子在结构上应该相同.具体做法是,将源语句中的某个词通过 BERT 遮蔽语言模型[26]进行替换,生成上下文相似的句子,最后再比较这两个句子的结构相似度.Zheng 等人[27]也提出了一种自动测试神经机器翻译系统的方法,通过短语识别和联系学习可以自动发现神经机器翻译系统的过译(over-translation)和漏译(under-translation)错误.Shashij 等人[28]提出了一种翻译错误的自动检测方法,借助句子的成分句法分析树将句子中的短语独立出来,通过比较短语在句子中和独立翻译的结果来自动发现系统的翻译错误.Sun 等人[29]提出了一个结合测试与修复的框架 TransRepair,在测试阶段也采用了基于词替换的方法来生成上下文相似句子.

但是,上述基于蜕变测试的工作仍然存在不足之处.MCMT [23]采用随机算法来选择中间语言,但不同语言的翻译效果有较大差异,会对实验产生干扰.MT4MT[24]设计的替换规则过于主观,能被替换的词的范围较小.另外,基于蜕变测试的已有工作都只采用了一种蜕变关系来进行蜕变测试,实验结果缺乏说服力.

针对已有研究工作存在的不足,本文提出了一个多粒度的蜕变测试框架 MGMT,可以在无需参考译文的情况下对神经机器翻译系统进行鲁棒性评估.MGMT 与已有方法有较大区别.首先,MGMT 与已有工作的目的不同.已有工作旨在利用蜕变测试思想对神经机器翻译系统的翻译性能进行评估(例如 RTT[22]、MCMT[23])或定位翻译错误的样本(例如 SIT[25]、TransRepair[29]),主要关注翻译的正确性;而 MGMT 的主要目的是借助蜕变测试对神经机器翻译系统的整体鲁棒性进行评估,主要关注翻译的稳定性.具体来说,MGMT 与 RTT[22]、MCMT[23]、MT4MT[24]都基于蜕变测试对神经机器翻译系统进行整体性评估且不需要参考译文,RTT[22]、MCMT[23]、MT4MT[24]基于单一蜕变关系来评估翻译质量,但 MGMT 旨在以合理的方式评估翻译系统鲁棒性,因此使用了 3 种符合翻译直觉的蜕变关系(具体细节可参考第 3.2 节中的蜕变关系定义).MGMT 中短语和单词粒度的测试样本生成思路受 SIT[25]和 TransRepair[29]的启发,用替换的方式生成测试样本,但 SIT[25]的目的是尽可能地发现更多的翻译错误,因此采用了尽可能多且独立于替换方法的相似度计算方法,而 MGMT 为确保鲁棒性评估的合理性,采用了一一对应的替换方法和相似度计算方法.

3 多粒度的蜕变测试框架

在本节中,首先介绍框架的整体架构和测试流程(见第 3.1 节).其次介绍框架中的主要模块设计,包括句子粒度、短语粒度、单词粒度上的蜕变关系定义(见第 3.2 节).之后介绍 MGMT 框架中如何选择待替换成分并进行成分替换(见第 3.3 节)以及如何进行各粒度上的相似度计算(见第 3.4 节).

3.1 整体架构

本文提出的面向神经机器翻译系统的多粒度蜕变测试框架 MGMT 大致可分为 3 个部分.图 3 展示了从源语句输入到蜕变关系判定结果输出的主要流程.

(1) 选择源语句中的待替换成分.根据 MGMT 中定义的蜕变关系(见第 3.2 节),对源语句进行句子、短语、单词粒度的替换.因此,首先要在 3 个粒度上选择源句子中需要替换的成分.在 MGMT 的设计中,源句子在句子粒度上的待替换成分即为整个句子.接着进行待替换单词和待替换短语的选择.我们对源语句进行成分句法分

© 中国科学院软件研究所 http://www.jos.org.cn

析,得到句子的成分句法分析树,再使用 DeepSelect 算法(见第 3.1.1 节)在成分句法分析树上进行选择.在图 3 所示的例子中,我们根据分析树选择了一个 NNP(proper noun,singular)词性单词作为待替换单词和一个 ADJP (adjective phrase)词性短语作为待替换短语.

(2) 对源语句进行成分替换.句子粒度上的替换基于 RTT[22]思想.RTT 包含正译(forward translation,简称 FT)和回译(backward translation,简称 BT),正译是指将文本从源语言翻译到目标语言,回译是指将正译得到的翻译结果翻译回源语言.我们先将源语句正译到目标语言,再回译到源语言以得到句子粒度的替换结果.短语粒度和单词粒度上的替换基于 BERT 遮蔽语言模型[26].本文将上面(1)中选中的待替换单词和短语用遮蔽词替代,再输入 BERT 遮蔽语言模型中,该模型可以根据句子的语境预测被遮蔽位置的词.最后用预测出的结果替换源句子中相同位置的单词和短语以得到短语粒度和单词粒度的替换结果.

(3) 翻译并对翻译结果进行相似度计算.将源语句连同上面(2)中 3 个粒度的替换语句输入到神经机器翻译系统中得到 4 个目标语言翻译结果,并分别对 3 个粒度上的替换语句和源语句的翻译结果进行相似度计算.在句子粒度上,根据编辑距离[12]分别计算源语言句子对和目标语言句子对的相似度.在短语和单词粒度上,考虑到选择待替换成分是基于成分结构分析树的,因此在计算目标语言句子对相似度时也基于句子的成分结构分析树.最后根据相似度计算结果判断是否满足 MGMT 定义的蜕变关系(见第 3.2 节).

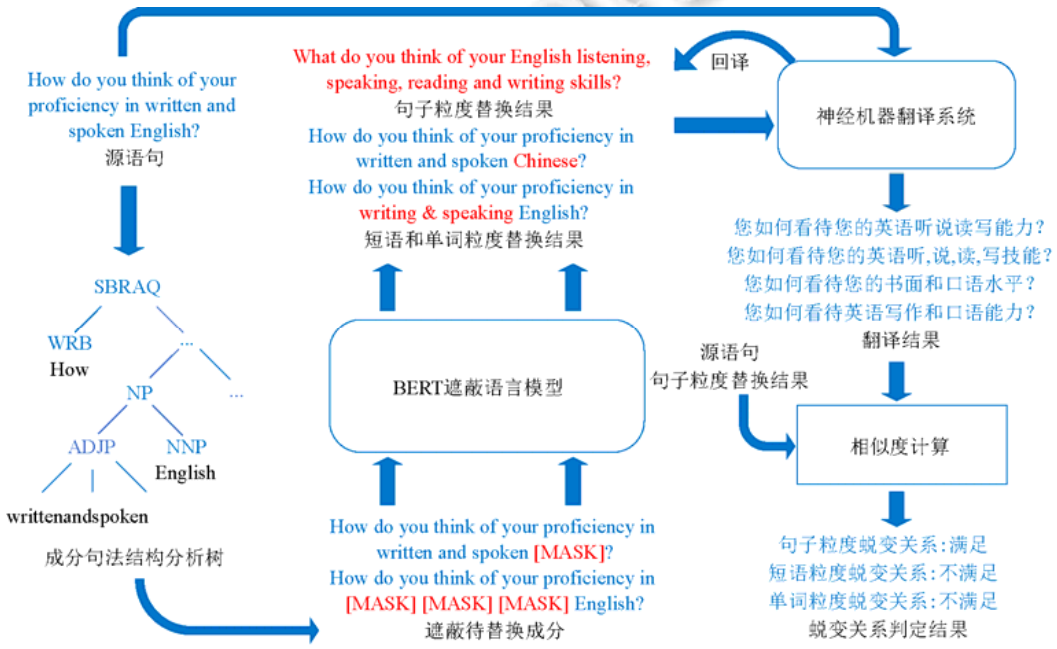

Fig.3 Process of multi-granularity metamorphic testing framework

图 3 多粒度蜕变测试框架的流程

3.2 蜕变关系定义

为了利用蜕变测试对神经机器翻译系统进行合理的翻译鲁棒性评估,在本节中我们定义了句子、短语和单词 3 个粒度的蜕变关系.这 3 个蜕变关系的定义基于对翻译任务的常理性推断.句子粒度的蜕变关系基于:源语句的直译结果与源语句经过多轮翻译得到的翻译结果应该接近.短语和单词粒度的蜕变关系基于:改变源语句中的一小部分,那么源语句其他部分的翻译结果应该不变.下文第 3.2.1 节、第 3.2.2 节和第 3.2.3 节分别详细介绍句子、短语、单词粒度的蜕变关系定义及判定方法.

3.2.1 句子粒度蜕变关系

RTT[22]是在没有参考译文情况下的一种常用机器翻译系统测试手段.其测试流程是首先将源语言正译成

目标语言,再将目标语言翻译结果回译到源语言,最后通过比较两个源语言句子来评估机器翻译系统的质量.本文在 RTT“正译-回译”流程的基础上添加 1 次正译,由此定义了句子粒度的蜕变关系 MR_{sl}.

定义 1(句子粒度蜕变关系 MR_{sl}). 设源语言句子为 S,将 S 经过神经机器翻译系统正译得到目标语言翻译结果 S_t,再将 S_t 通过神经机器翻译系统回译到源语言得到翻译结果 S_1,最后将 S_1 通过神经机器翻译系统再一次正译到目标语言得到 S_{t1},那么 S、S_1、S_t、S_{t1} 应满足:

$$Similarity(S_t,S_{t1})/Similarity(S,S_1)\geqslant 1 \tag{1}$$

公式(1)的含义是,用目标语言句子对与源语言句子对相似度的比值来评估基于句子粒度的翻译鲁棒性,目的是排除回译对实验的影响.MGMT 框架实际评估的是神经机器翻译系统在源语言到目标语言翻译(正译)上的翻译鲁棒性,然而句子粒度的替换过程(如图 3 所示)涉及到一次回译.回译采用的神经机器翻译模型是与正译采用的神经机器翻译模型相互独立的,因此在回译阶段产生的翻译错误会影响第 2 次正译.例如,在某次测试过程中,正译的翻译质量极高而回译的翻译质量很低,直接以 $Similarity(S,S_1)$或 $Similarity(S_t,S_{t1})$评估翻译质量都会导致评估值远高于真实值.因此,我们在公式(1)中用 $Similarity(S_t,S_{t1})/Similarity(S,S_1)$作为翻译质量的评估值,意在为低质量的回译过程增加一个补偿因子:如果某次回译过程翻译质量较差($Similarity(S,S_1)$较小),那么正译翻译质量分数应得到部分补偿(即 $Similarity(S_t,S_{t1})/Similarity(S,S_1)$的值会增大);若回译过程翻译质量较好(即 $Similarity(S,S_1)$接近 1),那么翻译鲁棒性的真实值也更接近目标语言句子对的相似度 $Similarity(S_t,S_{t1})$,而此时公式(1)中的评估值 $Similarity(S_t,S_{t1})/Similarity(S,S_1)$也更接近 $Similarity(S_t,S_{t1})$.

3.2.2 短语粒度蜕变关系

一个句子由单词构成,不同的单词能够组成不同的短语结构.以英文句子为例,短语结构可分为名词性短语(noun phrase,简称 NP)、动词性短语(verb phrase,简称 VP)、介词性短语(prepositional phrase,简称 PP)、副词性短语(adverb phrase)等.将源句子中某个短语结构替换为另一个近似的短语结构之后,源句子和替换后句子经过神经机器翻译系统翻译得到的翻译结果的结构应该相同.本文由此定义了短语粒度的蜕变关系 MR_{pl}.

定义 2(短语粒度蜕变关系 MR_{pl}). 设源语句为 S,替换 S 中的某个短语产生结构相似的替换语句 S_p.再将 S 和 S_p 通过神经机器翻译系统翻译到目标语言得到结果 S_t 和 S_{pt}.那么,S_t 和 S_t 应满足:

$$StructureSimilarity(S_t,S_{pt})=1 \tag{2}$$

公式(2)的含义是源语句 S 和短语替换语句 S_p 经过神经机器翻译系统的翻译结果 S_t 和 S_{pt} 在结构上应该相同.本文用基于成分句法分析树的相似度计算方法(见第 3.3.3 节)来计算 S_t 和 S_{pt} 的结构相似度.结构相似度的取值范围在 0~1 之间,取值为 0 时,说明两个句子的句法分析树结构完全不同,取值为 1 时,说明两个句子的句法分析树结构完全相同.

3.2.3 单词粒度蜕变关系

一个句子由单词构成,不同的单词有着不同的词性,处在不同的句子结构块中.将源句子中的某个单词替换为相同上下文的近似单词,那么源句子和替换后的句子的翻译结果在结构上应该相同.本文由此定义了单词粒度的蜕变关系 MR_{wl}.

定义 3(单词粒度蜕变关系 MR_{wl}). 设源语句为 S,替换 S 中某个单词产生结构相似的替换语句 S_w,再将 S 和 S_w 通过神经机器翻译系统翻译到目标语言得到结果 S_t 和 S_{wt}.那么,S_t 和 S_{wt} 应满足:

$$StructureSimilarity(S_t,S_{wt})=1 \tag{3}$$

公式(3)的含义是源语句 S 和单词替换语句 S_w 经过神经机器翻译系统得到的翻译结果 S_t 和 S_{wt} 在结构上应该相同.同样,我们用基于成分句法分析树的相似度计算方法(见第 3.3.3 节)来计算 S_t 和 S_{wt} 的结构相似度.结构相似度的取值范围为 0~1,取值为 0 时,说明两个句子的句法分析树结构完全不同,取值为 1 时,说明两个句子的句法分析树结构完全相同.

3.3 替 换

3.3.1 选择待替换成分

根据 MGMT 定义的蜕变关系(见第 3.2 节),对于一个测试样本要在 3 个粒度上进行基于替换的蜕变测试,

© 中国科学院软件研究所 http://www.jos.org.cn

因此首先要在句子、短语、单词 3 个粒度上选择源语句中要被替换的成分.

句子粒度的待替换成分即整个源句子.短语和单词粒度的待替换成分是句子中的某个短语和单词.在 MGMT 中,基于源句子的成分句法分析树来选择短语和单词粒度的待替换成分.图 4 展示了一个英文句子经过 BerkeleyParser[30]得到的成分句法分析树.可以看到,句子根据单词词性和短语词性被组织成树状结构.成分句法分析树的节点都是由句子中的单词构成的,每一棵子树都是某几个单词的组合.那么在句子中选择一个词或一个短语,就等价于在句法分析树中选择到达某棵子树的路径.基于以上特点,本文设计了一种基于成分句法分析树的选择算法 DeepSelect 来在短语粒度和单词粒度上选择要替换的成分.

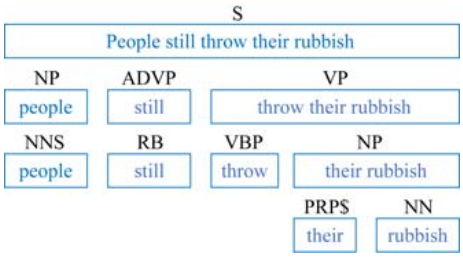

Fig.4 Example of constituency parse tree

图 4 成分句法分析树示例

算法 1. DeepSelect.

输入:源语句 *Sentence*,候选集大小 *Candidatenum*,短语词性集合 *PhrasePOS*;

输出:待替换路径 *finalpath*.

```
1.  建立待替换路径的候选集 Candidates,初始化为空集
2.  Tree=BerkeleyParser(Sentence)
3.  遍历 Tree,将 Tree 中的每一条路径(包含子路径)加入路径集合 PathSet 并将 PathSet 中的元素按路径长
    度降序排列
4.  if 当前粒度为单词粒度 then
5.   for each path in PathSet do
6.          if length(Candidates)<Candidatenum then
7.            Candidates.append(path)
8.  elif 当前粒度为短语粒度 then
9.   for each path in PathSet dopath in PathSet
10.         if length(Candidates)>Candidatenum and Tree[path] in phrasePOS then
11.           Candidates.append(path)
12. Finalpath=random.choose(Candidates)
13. return Finalpath
```

DeepSelect 算法旨在选择句子在短语和单词粒度下的待替换成分.首先,采用 BerkeleyParser[30]句法分析器来生成句子的成分句法分析树.由于处在成分句法分析树较深路径的节点的粒度一般较小,选择这些节点更符合 MGMT 的蜕变关系定义(见第 3.2 节),因此我们将句法分析树中的路径按从长到短排序收集到路径集合 *PathSet* 中.接着,在单词粒度下,我们直接往候选集中添加 *Candidatanum* 条路径;在短语粒度下还需进行一个额外判断,要求路径节点的词性必须是短语结构型.最后,为了保证路径选择的公平性,算法随机从候选集中选择一条路径作为最终的待替换路径.

3.3.2 成分替换

句子粒度的成分替换采用的是基于往返翻译的方法.首先将源句子输入神经机器翻译系统得到目标语言的直译结果,再将直译结果输入翻译系统翻译回源语言,这样就得到了一个句子粒度的替换语句.

短语粒度和单词粒度的成分替换采用的是 BERT[26]遮蔽语言模型.BERT 模型是一个非常成功的自然语言理解模型,在很多自然语言处理任务中通过微调都能达到 SOTA(state of the art)效果.在 BERT 中每个词的词向量不是唯一的,而是与词的上下文相关,因此,通过 BERT 可获得符合句子语义的词向量.模型主要通过遮蔽词预测和下一句预测这两个任务来进行训练.其中,遮蔽词预测是指将一个句子中 15%的词遮蔽,把预测这些被遮蔽

© 中国科学院软件研究所 http://www.jos.org.cn

位置的词当作目标任务来进行损失计算和模型参数优化.BERT 遮蔽语言模型是 BERT 的一部分,用遮蔽词任务进行训练,可以预测出句子中被遮蔽位置的词.

本文对短语粒度和单词粒度的成分替换采用一个预训练好的 BERT 遮蔽语言模型来实现,图 5 展示了预测遮蔽词的原理与流程.

(1) 将句子中待替换的部分置换为遮蔽词[MASK].

(2) 将遮蔽后的句子输入 BERT 遮蔽语言模型.句子经过 BERT 中的 Transformer 编码器转换成词向量,并输出每个位置上对于词典中每个词的预测分数.

(3) 对被遮蔽位置上的预测分数进行 argmax,得到概率最大的词(需与源句子中的待替换词不同).

(4) 将得到的替换词代替原词,得到替换后语句.

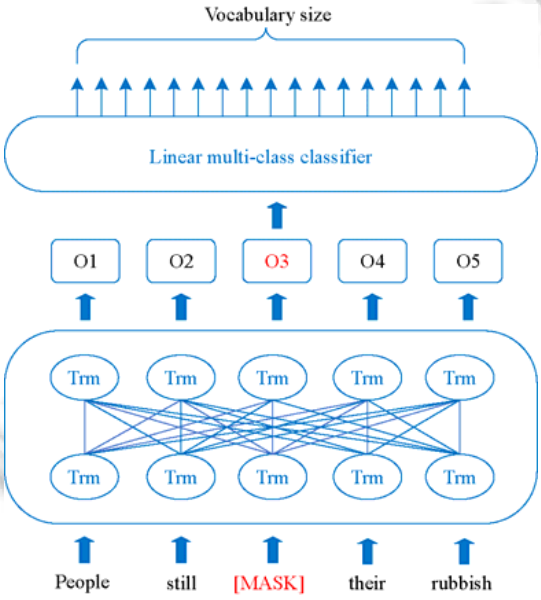

Fig.5 Process of predicting mased word by BERT maked language model

图 5 BERT 遮蔽语言模型预测遮蔽词的过程

3.3.3 句子粒度的相似度计算

句子粒度的蜕变测试涉及 3 次翻译和 4 个句子,需分别计算源语言和目标语言的句子对的相似度.

在本文中采用的是编辑距离[31]的方法,编辑距离属于文本相似度较常见的一种度量指标.通过计算从字符串 A 转换为字符串 B 最少需要操作的次数来衡量 A 和 B 的相似程度.编辑距离值越小,A 和 B 越相似.句子粒度的相似度计算公式为

$$Similarity(S_1,S_2)=1-\frac{2\times EditDistance(S_1,S_2)}{length(S_1)+length(S_2)} \tag{4}$$

在公式(4)中,S_1 和 S_2 是需要计算相似度的两个句子.*EditDistance* 计算的是两个句子间的编辑距离,实验中采用的是 NLTK(http://www.nltk.org/)实现的算法.*length* 函数计算的是句子长度.相似度取值在 0~1 之间.

3.3.4 短语粒度和单词粒度的相似度计算

对短语粒度和单词粒度的蜕变测试是基于源语句的成分句法分析树进行替换的,因此,这两个粒度的结构相似度计算也基于翻译结果的成分句法分析树.具体定义如下:

$$StructureSimilarity(S_1,S_2)=1-\frac{LostRate(S_1,S_2)+AddRate(S_1,S_2)}{2} \tag{5}$$

© 中国科学院软件研究所 http://www.jos.org.cn

$$LostRate(S_1,S_2)=\frac{PathSet(Tree_1)-PathSet(Tree_2)}{PathSet(Tree_1)} \tag{6}$$

$$AddRate(S_1,S_2)=\frac{PathSet(Tree_2)-PathSet(Tree_1)}{PathSet(Tree_2)} \tag{7}$$

在公式(5)~公式(7)中,S_1 和 S_2 为要计算结构相似度的两个句子.$Tree_1$ 和 $Tree_2$ 分别是 S_1 和 S_2 的成分句法分析树.*PathSet* 是句法分析树的所有路径(包含子路径)的集合.S_1 和 S_2 两个句子的结构相似度 *StructureSimilarity* 在数值上定义为丢失率 *LostRate* 和新增率 *AddRate* 的平均值的补数.*LostRate* 计算的是 $Tree_1$ 相对于 $Tree_2$ 中的路径丢失率,*AddRate* 计算的是 $Tree_2$ 相对于 $Tree_1$ 的路径新增率.结构相似度的取值在 0~1 之间.

4 实 验

为了验证本文提出的多粒度蜕变测试框架 MGMT 的有效性,我们尝试回答以下 3 个问题.

问题 1:MGMT 能否在无需参考译文的情况下对神经机器翻译系统进行质量评估?

问题 2:MGMT 质量评估的结果是否具有合理性和说服力?

问题 3:MGMT 与其他同类型的蜕变测试方法相比是否具有优越性?

实验问题 1 的目的是检验 MGMT 的可用性,需要基于 MGMT 框架在无参考译文的情况下对神经机器翻译系统进行质量评估,并给出量化的评估结果.实验问题 2 的目的是检验 MGMT 实验结果的合理性,需要将 MGMT 实验结果与基于参考译文的质量评估结果进行比较.实验问题 3 的目的是比较 MGMT 与同类型方法的评估准确度,需要将 MGMT 实验结果与同类型的蜕变测试实验结果进行比较.

4.1 实验设计

为了回答上述 3 个实验问题,本文进行了如下实验设计.

首先是数据集和实验对象选择.我们选择一个公开的中英对照数据集 UM-Corpus[7]作为数据来源,选取了其中 5 个领域(教育、微博、新闻、口语、字幕),每个领域数据集中的前 1 000 个中英句子对,总共 5 000 个句子对作为实验数据集.然后,本文选择了 6 个使用广泛、多语种覆盖的产业界神经机器翻译系统(阿里翻译、百度翻译、必应翻译、谷歌翻译、腾讯翻译、搜狗翻译)作为质量评估对象.

针对问题 1,本文将实验数据集中的 5 000 个英文句子作为源语句,使用本文提出的多粒度蜕变测试框架 MGMT 对 6 个神经机器翻译系统在 5 个领域下的翻译鲁棒性进行了评估,记录了各神经机器翻译系统在句子粒度、短语粒度和单词粒度的蜕变关系满足率,将用于评估的句子集在 3 个粒度上的蜕变关系满足率平均值作为神经机器翻译系统的最终翻译鲁棒性分数.由于我们提出的用于短语粒度和单词粒度蜕变替换的 DeepSelect 算法(见第 3.3.1 节)具有随机性,总共进行了 3 次翻译鲁棒性分数测定实验,采用 3 次实验数据的平均值作为最终实验结果,以确保实验结果的可靠性.

针对问题 2,需要检验 MGMT 的测试结果是否符合实际.机器翻译系统的翻译鲁棒性没有专门的评估指标,在过往工作中通常是根据翻译质量来加以衡量.基于参考译文计算文本相似度是翻译质量评估的最常见手段,大部分翻译质量评估任务采用的都是这种方法.基于参考译文的方法具有较高的合理性和说服力,因此要证明 MGMT 的合理性,只需将其实验结果与基于参考译文的文本相似度计算结果进行比较.由此,本文选择了 3 个使用广泛且具有代表性的文本相似度指标:基于编辑距离的 WER[17]、基于精确率的 BLEU[15]以及基于召回率的 METEOR[16].本文利用这 3 个相似度指标设置了一个基于参考译文的参照实验,选择实验数据集中的 5 000 个中文句子作为参考译文,与 5 000 个英文语句通过神经机器翻译系统得到的直译结果作比较,计算出各神经机器翻译在各领域数据集上翻译结果的 BLEU、METEOR、WER 数值,与 MGMT 的实验结果进行比较.

针对问题 3,需要将 MGMT 的实验结果与同类型方法作比较.用蜕变测试对神经机器翻译系统进行质量评估的方法主要有 RTT[22]、MCMT[23]及 MT4MT[24]这 3 种,由于 MT4MT 的实验方法只适用于特定类型的数据,因此本文选择了 RTT 和 MCMT 作为对比方法进行实验,将 RTT、MCMT、MGMT 对各神经机器翻译系统的评估结果与实验问题 2 中基于参考译文的 BLEU、METEOR、WER 数值进行比较.与基于参考译文的评估结

© 中国科学院软件研究所 http://www.jos.org.cn

果越接近,说明方法的准确性更高.但是,各组实验结果的量纲不同,对实验结果进行绝对数值上的比较不具备一定的参考意义,因此,本文选择皮尔逊相关系数(Pearson correlation coefficient,简称 PC)[32]和斯皮尔曼等级相关系数(Spearman rank correlation,简称 SRC)[33]作为实验结果相似度的评测指标.

4.2 实验结果与分析

4.2.1 针对问题1的结果分析

针对问题 1,本文基于 MGMT 开展了大规模的实证研究,在 5 个领域的数据集上对 6 个神经机器翻译系统进行了翻译鲁棒性评估.对于每个领域数据集中的 1 000 个英文句子,MGMT 生成了句子、短语、单词 3 个粒度上共 3 000 个替换语句,输入神经机器翻译系统得到 4 000 个目标语言翻译结果,并根据源语句和翻译结果进行相似度计算和蜕变关系判定.我们用蜕变关系满足率作为系统翻译鲁棒性的量化指标.对于每个测试样本,MGMT 会基于第 3.2 节中定义的 3 个蜕变关系进行句子、短语、单词粒度上的蜕变测试.当测试样本满足某个粒度的蜕变关系时,该粒度下的蜕变关系满足率记为 1,否则,记为 0.测试样本违背某个粒度的蜕变关系,说明神经机器翻译系统对于测试样本在该粒度上的翻译鲁棒性较差.我们用各神经机器翻译系统在 3 个粒度上的平均蜕变关系满足率作为最终的系统质量评估分数,分数越高,说明该系统的整体翻译鲁棒性越好.

表 1 展示了基于 MGMT 框架对 6 个神经机器翻译系统(阿里翻译、百度翻译、必应翻译、谷歌翻译、腾讯翻译、搜狗翻译)在 5 个领域(教育、微博、新闻、口语、字幕)的中英文翻译数据集上测得的翻译鲁棒性分数.表中加粗数据代表同一领域不同神经机器翻译系统的最高质量分数或同一神经机器翻译系统在不同领域中的最高质量分数.根据表 1,可以分析得到以下结论.

(1) 各神经机器翻译系统存在鲁棒性差异.阿里翻译系统在微博、新闻和字幕 3 个领域数据集中的质量分数都排名第 1,在口语和教育领域排名第 2,总体鲁棒性最好.必应翻译和谷歌翻译在 5 个领域数据集上的排名都在第 5 位和第 6 位,与其他神经机器翻译系统在质量上有显著差异,总体鲁棒性最差.百度翻译、腾讯翻译、搜狗翻译在各领域的鲁棒性差异不大.

(2) 不同领域数据上系统鲁棒性存在差异.在微博领域,各神经机器翻译系统表现出的鲁棒性最好,平均质量分数在 5 个领域中最高.而在新闻领域的质量分数较低,质量分数平均值没有超过 40,说明各神经机器翻译系统在新闻领域的鲁棒性较差.

综合各领域数据集上的评估结果,可以得出在 MGMT 方法下各神经机器翻译系统翻译鲁棒性排名如下:

1.阿里翻译 2.百度翻译 3.搜狗翻译 4.腾讯翻译 5.必应翻译 6.谷歌翻译

以上的实验结果和结论可以回答问题 1,说明本文提出的评估框架 MGMT 无需参考译文即可对神经机器翻译系统进行鲁棒性评估.

4.2.2 针对问题2的结果分析

针对问题 2,本文使用实验数据集中的 5 000 个中文句子作为参考译文,以 BLEU[15]、METEOR[16]、WER[17]作为相似度度量指标,计算各神经机器翻译系统的质量分数.另外,BLEU、METEOR 数值与句子的翻译质量成正比,而 WER 值与翻译质量成反比,因此,为了直观比较实验数据,实验中用于比较的 WER 数值为实际 WER 数值的补数.

为了进一步证明 MGMT 评估结果的合理性,本文对每个源语句的评估结果进行了更加具体的统计.在 MGMT 实验中,对每个源语言句子进行句子、短语和单词 3 个粒度上的蜕变测试,并根据相似度计算结果判定是否满足 3 个粒度的蜕变关系.为了分析句子层面的评估结果,本文将每个句子的蜕变关系满足率分为 4 个等级.0 代表该句子的翻译结果无法满足任何粒度的蜕变关系,1/3 代表满足 1 个粒度的蜕变关系,2/3 代表满足 2 个粒度的蜕变关系,1 代表满足所有粒度的蜕变关系.据此将所有句子按评估结果的蜕变关系满足率等级分为 4 组,并计算每组句子基于参考译文的 BLEU、METEOR、WER 数值,与蜕变关系满足等级进行比较.比较结果见表 2.由表 2 所示数据可知,在教育、微博、新闻、口语、字幕这 5 个领域的数据集上,中英文句子对的平均 BLEU、METEOR、WER 数值是随着句子的蜕变关系满足率等级的提升而逐级提高的.也就是说,对于一个源语句,在 MGMT 下测得的蜕变关系满足率和基于句子参考译文计算出来的 BLEU、WER、METEOR 数值

© 中国科学院软件研究所 http://www.jos.org.cn

呈正相关性.真实翻译分数越低的句子对,在 MGMT 下测得的蜕变关系满足率也就越低.相关工作——SIT[25]和 TransRepair[29]利用蜕变测试来发现翻译错误,而 MGMT 除了衡量神经机器翻译系统的整体鲁棒性之外,也可用于发现翻译错误的样本.在实际操作中可以根据样本的蜕变关系满足率来加以判断,一个样本的蜕变关系满足率越低,那么它是一个翻译错误的可能性就越大(由于实验数据集过大,我们将在下一步工作中采用人工验证的方式对上述操作的可行性进行验证).

Table 1 Quality evaluation results of multi-granularity metamorphic testing framework

表 1 多粒度测试框架质量评估结果

神经机器翻译系统	蜕变关系粒度	数据集领域				
		教育	微博	新闻	口语	字幕
阿里翻译	句子粒度	46.5	48.6	39.8	41.8	47.6
	短语粒度	26.0	35.8	24.6	34.0	34.0
	单词粒度	38.4	44.0	41.2	46.0	43.8
	平均值	36.9	**42.8**	**35.2**	40.6	**41.8**
百度翻译	句子粒度	51.0	52.0	45.0	49.4	48.0
	短语粒度	25.6	31.6	23.2	30.0	28.8
	单词粒度	35.4	41.2	36.0	44.2	38.6
	平均值	37.3	**41.6**	34.7	41.2	38.5
必应翻译	句子粒度	38.6	40.2	33.0	36.0	39.4
	短语粒度	17.8	22.4	15.8	22.8	18.6
	单词粒度	32.0	31.2	36.6	42.2	35.8
	平均值	29.5	31.2	28.4	**33.6**	31.2
谷歌翻译	句子粒度	43.2	37.8	39.4	41.2	43.2
	短语粒度	9.0	18.2	10.4	14.6	19.0
	单词粒度	22.0	30.4	24.2	30.0	30.0
	平均值	24.7	28.8	24.7	28.6	**30.7**
腾讯翻译	句子粒度	46.2	48.8	34.0	42.4	44.8
	短语粒度	27.1	35.0	24.0	30.0	25.2
	单词粒度	38.8	38.6	44.0	46.6	40.8
	平均值	37.4	**40.8**	34	39.7	36.9
搜狗翻译	句子粒度	49.0	48.0	41.4	48.4	53.0
	短语粒度	24.2	30.0	17.0	29.8	23.0
	单词粒度	40.6	41.2	38.8	49.6	38.8
	平均值	**37.9**	39.7	32.4	**42.6**	38.2

Table 2 Comparison of evaluation results base on reference translations and evaluation results based on MGMT (group by satisfaction rate of metamorphic relationships)

表 2 MGMT 评估结果(按蜕变关系满足率分组)与基于参考译文的评估结果比较

相似度指标	蜕变关系满足等级	中英文数据集领域					
		教育	微博	新闻	口语	字幕	平均
BLEU	0	18.4	20.3	18.9	23.2	17.1	20.0
	1/3	25.3	26.2	23.3	26.0	22.1	24.9
	2/3	29.5	28.5	28.5	26.7	23.7	27.2
	1	**38.2**	**33.6**	**33.6**	**30.1**	**29.5**	**31.8**
METEOR	0	40.1	41.2	41.6	46.1	33.2	41.7
	1/3	49.2	49.0	47.1	49.4	42.0	48.1
	2/3	53.9	52.0	49.8	50.4	43.6	50.9
	1	**61.4**	**56.8**	**50.3**	**53.5**	**52.2**	**55.3**
WER	0	25.8	30.3	24.9	34.7	19.4	28.1
	1/3	37.8	40.4	32.6	38.6	29.9	36.7
	2/3	43.2	43.9	**35.6**	40.2	31.7	40.0
	1	**54.1**	**50.0**	35.3	**45.5**	**42.1**	**46.1**

综上所述,MGMT 的评估结果与基于参考译文的 BLEU、METEOR、WER 的评估结果相似度较高,可以证明,MGMT 对各神经机器翻译系统的评估结果具有合理性和说服力.

4.2.3 针对问题3的结果分析

针对问题 3,本文用评测指标 PC 和 SRC,将基于蜕变测试的方法 RTT[22]、MCMT[23]、MGMT 与基于参考

© 中国科学院软件研究所 http://www.jos.org.cn

译文的基准指标 BLEU[15]、METEOR[16]、WER[17]进行比较.PC 和 SRC 衡量的是两组数据在变化方向和数据排名上的相关度,取值均在–1 到 1 之间,–1 代表完全负相关,1 代表完全正相关.实验中的 PC 值和 SRC 值通过 Python 语言的 Scipy[34]包进行计算.

在进行 RTT 实验时,本文完全参照 Milam 等人[22]的方法,先经过 FT(forward translation)得到目标语言翻译结果,再将目标语言翻译结果通过 BT(backward translation)得到源语言翻译结果,最后用 BLEU 指标计算源语言翻译结果与源语句的相似度,并以此作为各神经机器翻译系统的质量分数.在进行 MCMT 实验时,我们在 Daniel[23]等人方法的基础上进行了微小变动.原方法中 MCMT 在 7 种语言(法语、日语、韩语、西班牙语、俄语、葡萄牙语、瑞典语)中随机选一个作为中间语言,但本文实验中并不是所有神经机器翻译系统都支持瑞典语的翻译,所以在本文复现的 MCMT 实验中,中间语言选择范围调整为 6 种语言(法语、日语、韩语、西班牙语、俄语、葡萄牙语).

我们将基于蜕变测试的 RTT、MCMT、MGMT 这 3 种方法测得的质量分数与第 4.2.2 节中实验得到的基于参考译文的 BLEU、METEOR、WER 数值按照神经机器翻译系统进行分组,每组数据包含该神经机器翻译系统在 5 个领域上测得的质量分数.之后我们计算基于蜕变测试方法的每一组数据与基于参考译文的质量分数的 PC 值和 SRC 值.PC 值、SRC 值及相应的 p 值(p-value)的计算结果见表 3 和表 4.

Table 3 Similarity comparision of evaluation results based on metamorphic testing and reference translations (PC)

表 3 基于蜕变测试与基于参考译文的实验结果相似度比较(PC)

基准指标	实验方法	教育		微博		新闻		口语		字幕		平均	
		相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值
BLEU	RTT	0.66	0.16	0.53	0.28	0.27	0.60	0.56	0.25	0.27	0.60	0.46	0.38
	MCMT	0.61	0.20	–0.02	0.96	0.20	0.70	0.52	0.29	0.29	0.58	0.32	0.55
	MGMT	**0.94**	**0.01**	**0.91**	**0.01**	**0.78**	**0.06**	**0.90**	**0.01**	**0.67**	**0.14**	**0.84**	**0.05**
METEOR	RTT	0.67	0.14	0.48	0.34	0.31	0.27	0.54	0.27	0.20	0.70	0.44	0.40
	MCMT	0.53	0.28	0.01	0.98	0.20	0.71	0.55	0.26	0.10	0.84	0.28	0.61
	MGMT	**0.92**	**0.01**	**0.93**	**0.01**	**0.83**	**0.04**	**0.91**	**0.01**	**0.65**	**0.16**	**0.85**	**0.05**
WER	RTT	0.65	0.17	0.57	0.24	0.29	0.58	0.58	0.23	0.21	0.68	0.46	0.38
	MCMT	0.60	0.20	0.05	0.91	0.26	0.62	0.56	0.24	0.20	0.70	0.34	0.54
	MGMT	**0.94**	**0.01**	**0.91**	**0.01**	**0.81**	**0.05**	**0.92**	**0.01**	**0.65**	**0.16**	**0.85**	**0.05**

Table 4 Similarity comparision of evaluation results based on metamorphic testing and reference translations (SRC)

表 4 基于蜕变测试与基于参考译文的实验结果相似度比较(SRC)

基准指标	实验方法	教育		微博		新闻		口语		字幕		平均	
		相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值	相关系数	p 值
BLEU	RTT	0.66	0.16	**0.66**	**0.16**	0.26	0.62	0.71	0.11	0.38	0.46	0.53	0.30
	MCMT	0.26	0.62	0.14	0.79	0.09	0.87	0.49	0.33	**0.6**	**0.20**	0.31	0.56
	MGMT	**0.94**	**0.01**	0.49	0.33	**0.43**	**0.40**	**0.83**	**0.04**	0.49	0.33	**0.63**	**0.22**
METEOR	RTT	0.66	0.16	0.43	0.40	0.43	0.40	0.71	0.11	0.20	0.70	0.49	0.35
	MCMT	0.26	0.62	0.09	0.87	0.2	0.70	0.49	0.33	0.31	0.54	0.27	0.61
	MGMT	**0.94**	**0.01**	**0.6**	**0.21**	**0.49**	**0.33**	**0.83**	**0.04**	**0.43**	**0.40**	**0.65**	**0.20**
WER	RTT	0.66	0.16	**0.66**	**0.16**	0.26	0.62	0.71	0.11	0.38	0.46	0.53	0.30
	MCMT	0.26	0.62	0.14	0.79	0.09	0.87	0.49	0.33	**0.60**	**0.21**	0.31	0.56
	MGMT	**0.94**	**0.01**	0.49	0.33	**0.43**	**0.40**	**0.83**	**0.04**	0.49	0.33	**0.63**	**0.22**

PC 指标反映的是两组实验结果在数据变化方向上的相关程度.从表 3 可以看出,在 BLEU、METEOR 和 WER 这 3 个基准指标上,MGMT 的 PC 值在各领域都显著高于 RTT 和 MCMT.MGMT 与 BLEU 指标平均相关系数为 0.85,p 值为 0.05;与 METEOR 和 WER 指标的相关系数平均值为 0.85,p 值为 0.05.从各领域平均值来看,MGMT 与 3 个基准指标的 PC 值都显著高于 RTT 和 MCMT(比 RTT 高约 83%,比 MCMT 高约 130%),且 p 值较低,说明 PC 值较为可信.以上数据说明,MGMT 的实验结果在数据变化方向上与 3 个基准指标的相关性更高,更接近基于参考译文的方法.

© 中国科学院软件研究所 http://www.jos.org.cn

SRC 指标反映的是两组实验结果数值在数据集中排名的相关程度.从表 4 可以看出,在基准指标 BLEU 和 WER 上,MGMT 的 SRC 值在教育、新闻、口语 3 个领域上最高,在微博领域上低于 RTT,在字幕领域上低于 MCMT.在基准指标 METEOR 上,MGMT 的 SRC 值在 5 个领域都达到最高.从各领域平均值来看,MGMT 与 3 个基准指标的 SRC 值都略高于 RTT(高约 20%),显著高于 MCMT(高约 100%),说明 MGMT 的实验结果在数据排名相关性上更接近基于参考译文的方法.

综上所述,MGMT 与 RTT、MCMT 相比,在两个相关系数 PC 和 SRC 上都更高.从 PC 值来看,MGMT 在 5 个领域数据集上的 PC 值都显著高于 RTT 和 MCMT,从 SRC 值来看,MGMT 在微博领域低于 RTT,在字幕领域低于 MCMT,但 5 个领域上的平均 SRC 值为最高.说明无论从数值角度还是将数值转化为排名后比较各神经机器翻译系统的质量,MGMT 的评估结果都更接近基于参考译文的方法,评估准确度比 RTT 和 MCMT 更高.

4.3 扩展讨论

4.3.1 蜕变粒度之间的关系

首先我们对蜕变粒度之间是否存在相关性进行研究.从表 5 可以看出,子粒度与单词粒度和短语粒度的 SRC 值都很低,分别为 0.027 和 0.041,说明句子粒度的判定结果与短语、单词粒度都不具有相关性.而短语粒度和单词粒度的 SRC 值为 0.274,说明这两个粒度的判定结果具有一定的相关性.短语粒度和单词粒度具有相关性是由于短语粒度和单词粒度都基于依存句法分析树和BERT模型进行蜕变测试.而句子粒度的蜕变测试流程相对独立.

在第 4.2 节中,我们证明了 MGMT 方法能够在没有参考译文的情况下对神经机器翻译系统进行质量评估,且优于同类型的 RTT 和 MCMT 方法.MGMT 方法的关键在于单词、短语和句子 3 个粒度的蜕变关系.为了探究哪个粒度的判定结果更具重要性,我们进行了消融实验.首先将句子根据满足的蜕变关系进行分类,计算出每一组句子的真实 BLEU、METEOR、WER 值,并采用控制变量的方式进行比较.如表 6 所示,不满足任何一个蜕变关系的句子的平均 BLEU 值为 0.178 9,而只满足单词粒度蜕变关系的句子的平均 BLEU 值为 0.190 7,相对质量分数提升了 0.011 8,看作单词粒度对真实质量分数的贡献值.在 BLEU、METEOR 和 WER 这 3 个基准指标上,句子粒度带来的平均增加值分别为 0.118 3、0.150 4、0.196 7,都远大于单词粒度(0.014 8、0.017 8、0.020 2)和短语粒度(0.023 6、0.026 2、0.033),说明与单词粒度和短语粒度的蜕变测试相比,句子粒度蜕变测试对句子真实翻译质量的影响更大.我们由此得出结论:在对真实质量分数的影响程度上,句子粒度的蜕变测试最高,其次是短语粒度,最后是单词粒度.

4.3.2 MGMT准确性原理分析

由第 4.2.2 节和第 4.2.3 节我们得出结论:用 MGMT 测量神经机器翻译系统鲁棒性具有一定的合理性和说服力,MGMT 与基于参考译文的方法正相关性较高,且显著优于同类型的 RTT 和 MCMT 方法.我们尝试分析其中的原因.首先,RTT 中涉及了一次正译和一次回译,正议和回译是独立的两个翻译过程(可以看作两个独立的神经机器翻译系统).回译过程可能影响机器翻译系统质量的测定;在 MGMT 的句子粒度蜕变关系中也涉及回译,为了降低回译过程中的翻译错误对整体鲁棒性测量带来的影响,我们设计了基于相对相似度的蜕变关系:当回译质量过低时,正译的质量分数应该得到一定的补偿.而 MCMT 通过中间语言来构造蜕变关系,但不同语言间的翻译差距也会影响机器翻译系统质量的测定.MGMT 方法在句子粒度蜕变关系设计过程中对非正译过程的影响进行了补偿处理,并结合单词粒度和短语粒度进一步提高了评估准确性,因此优于 RTT 和 MCMT.

Table 5 Comparison of the results of different metamorphic relationships

表 5 不同蜕变关系判定结果相关性比较

	SRC 值	*p* 值
句子粒度-字符粒度	0.027	0.000 8
句子粒度-短语粒度	0.041	6.8138e-07
短语粒度-字符粒度	0.274	3.5209e-256

© 中国科学院软件研究所 http://www.jos.org.cn

Table 6 The contribution of different transformation relations to the quality of authentic translation

表 6 不同蜕变关系对真实翻译质量的贡献值

	蜕变关系满足情况			平均	质量分数增量		
	单词	短语	句子	质量分数	单词	短语	句子
BLEU				0.178 9	–	–	–
	✓			0.190 7	0.011 8	–	–
		✓		0.200 4	–	0.021 5	\
			✓	0.285 8	–	–	0.106 9
	✓	✓		0.202 8	0.002 4	0.012 1	–
		✓	✓	0.323 1	–	0.037 3	0.122 7
	✓		✓	0.315 9	0.030 1	–	0.125 2
平均值					0.014 8	0.023 6	**0.118 3**
METEOR				0.388 1	–	–	–
	✓			0.401 7	0.013 6	–	–
		✓		0.411 7	–	0.023 6	\
			✓	0.527 1	–	–	0.139
	✓	✓		0.418 3	0.006 6	0.016 6	–
		✓	✓	0.565 6	–	0.038 5	0.153 9
	✓		✓	0.560 2	0.033 1	–	0.158 5
平均值					0.017 8	0.026 2	**0.150 4**
WER				0.245 2	–	–	–
	✓			0.267 4	0.022 2	–	–
		✓		0.278 2	–	0.033	–
			✓	0.428 4	–	–	0.183 2
	✓	✓		0.277 2	–0.001	0.009 8	–
		✓	✓	0.484 6	–	0.056 2	0.206 4
	✓		✓	0.467 8	0.039 4	–	0.200 4
平均值					0.020 2	0.033	**0.196 7**

4.4 有效性影响因素分析

本节我们来分析有可能影响实验有效性的影响因素.

(1) 内部有效性主要涉及影响实验结果正确性的内部因素.本文中的内部有效性影响因素是句法分析器的性能.我们使用的英文和中文句法分析器是 BerkeleyParser,该句法分析器在 WSJ 测试集上的 $F1$ 值可达 95.17,在 CTB5.1 测试集上的 $F1$ 值可达 91.69.可以将因句法分析错误导致的实验影响降到最小.

(2) 外部有效性主要涉及实验结果是否具有一般性.本文选择了机器翻译常用的一个公开中英文数据集 UM-corpus,并选取了其 5 个领域(包括教育、微博、新闻、口语和字幕)的翻译数据,翻译数据覆盖的领域较为全面,具有代表性.因此,基于该数据集的实验结果也具有可靠性和代表性.

(3) 结论有效性主要涉及评测指标的选择是否合理.本文为了评估基于蜕变测试和基于参考译文的实验结果的相似度,选择了两个相关系数指标——PC 和 SRC.PC 计算的是两组实验结果在数据变化方向上的相似度,SRC 计算的是数据排名的相似度.由于基于蜕变测试和基于参考译文的质量分数计算方法量纲上不同,基于绝对数值的比较不具备意义,因此,使用 PC 和 SRC 作为评测指标可以保证评估的合理性.另一个影响因素是文本相似度度量指标的选择是否合理.在实验中我们选择了 BLEU、METEOR、WER 作为文本相似度度量指标,这 3 个度量指标都被广泛使用且评估原理不同(BLEU 侧重精确率,METEOR 侧重召回率,WER 基于编辑距离),因此可以保证文本相似度度量的合理性.最后,在句子粒度蜕变关系(见 3.2.1 节)定义中涉及到了不同语言的相似度比较.相似度分布可能因不同的语言特性产生差异,这种差异无法避免.因此,我们对同一语言使用相同的分词工具,再将分词后的句子对基于编辑距离计算相似度,从而能够有效降低相似度分布差异对实验带来的影响.

5 总结与展望

基于神经网络结构的神经机器翻译系统应用广泛,许多翻译服务提供商的翻译服务都基于神经机器翻译系统,对其所进行的测试和质量评估也具有较高的研究和现实意义.本文提出了一个多粒度的蜕变测试框架 MGMT.该测试框架能够在没有参考译文的情况下对神经机器翻译系统进行质量评估.MGMT 首次使用了多粒

© 中国科学院软件研究所 http://www.jos.org.cn

度的蜕变关系对神经机器翻译系统进行整体性的翻译鲁棒性评估.从实验结果来看,MGMT 与已有的同类型方法相比,与基于参考译文的 BLEU、METEOR、WER 评估结果都更接近,评估准确度更高.

未来工作可以基于如下 3 个方面展开.首先,研究人员可以继续改进和优化本测试框架中的各个流程以达到更合理的质量评估效果,例如在替换阶段可以针对替换模型的缺陷设计有针对性地替换约束.其次,研究人员可以更改 MGMT 的部分设计,将其与错误定位任务结合起来,用于自动判定神经机器翻译系统的翻译错误.最后,我们将对本文中的实验数据和结果进行清理并共享,以方便研究人员针对翻译系统的鲁棒性展开后续研究.

References:

[1] Brown PF, Cocke J, Pietra SD, Pietra VJD, Jelinek F, Lafferty JD, Mercer RL, Roossin PS. A statistical approach to machine translation. Computational Linguistics, 1990,16(2):79–85.

[2] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the NIPS. 2014. 3104–3112.

[3] Li YC, Xiong DY, Zhang M. A survey of neural machine translation. Chinese Journal of Computers, 2018,41(12):100–121 (in Chinese with English abstract).

[4] Koehn P. Rebecca knowles: Six challenges for neural machine translation. In: Proc. of the NMT@ACL. 2017. 28–39.

[5] Cheng Y, Tu ZP, Meng FD, *et al*. Towards robust neural machine translation. ACL, 2018(1):1756–1766.

[6] Ma L, Xu FJ, Xue MH, Hu QG, *et al*. Simon see: Secure deep learning engineering: A software quality assurance perspective. CoRR abs/1810.04538, 2018.

[7] Tian L, Wong DF, Chao LS, Quaresma P, Oliveira F, Yi L. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. In: Proc. of the LREC. 2014. 1837–1842.

[8] https://translate.google.cn

[9] https://cn.bing.com/translator/

[10] https://fanyi.baidu.com

[11] https://translate.alibaba.com

[12] https://fanyi.qq.com

[13] https://fanyi.sogou.com/

[14] Chatzikoumi E. How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, 2020,26(2):137–161.

[15] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the ACL. 2002. 311–318.

[16] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. of the IEEvaluation@ACL. 2005. 65–72.

[17] Nießen S, Och FJ, Ney GLH. An evaluation tool for machine translation: Fast evaluation for MT research. In: Proc. of the LREC. 2000.

[18] Barr ET, Harman M, McMinn P, Shahbaz M, Yoo S. The Oracle Problem in software testing: A survey. IEEE Trans. on Software Engineering, 2015,41(5):507–525.

[19] Wang Z, Yan M, Liu S, Chen JJ, Zhang DD, Wu Z, Chen X. Survey on testing of deep neural networks. Ruan Jian Xue Bao/Journal of Software, 2020,31(5):1255−1275 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5951.htm [doi: 10.13328/j.cnki.jos.005951]

[20] McKeeman WM. Differential testing for software. Digital Technical Journal, 1998,10(1):100−107.

[21] Segura S, Fraser G, Sánchez AB, Cortés AR. A survey on metamorphic testing. IEEE Trans. on Software Engineering, 2016,42(9): 805–824.

[22] Aiken M, Park M. The efficacy of round-trip translation for MT evaluation. Translation Journal, 2010,14:1.

[23] Pesu D, Zhou ZQ, Zhen JF. Dave Towey: A Monte Carlo method for metamorphic testing of machine translation services. In: Proc. of the MET@ICSE. 2018. 38–45.

[24] Zhou ZQ, Sun LQ. Metamorphic testing for machine translations: MT4MT. In: Proc. of the ASWEC. 2018. 96–100.

[25] He PJ, Meister C, Su ZD. Structure-invariant testing for machine translation. In: Proc. of the ICSE. 2020. 961–973.

© 中国科学院软件研究所 http://www.jos.org.cn

[26] Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of Deep bidirectional transformers for language understanding. NAACL-HLT, 2019(1):4171–4186.

[27] Zheng WJ, Wang WY, Liu D, *et al*. Testing untestable neural machine translation: An industrial case. In: Proc. of the ICSE (Companion Volume). 2019. 314–315.

[28] Shashij G, He PJ, Clara M, *et al*. Machine translation testing via pathological invariance. In: Proc. of the ESEC/SIGSOFT FSE. 2020. 863–875.

[29] Sun ZY, Zhang JM, Harman M, *et al*. Automatic testing and improvement of machine translation. In: Proc. of the ICSE. 2020. 974–985.

[30] Kitaev N, Klein D. Constituency parsing with a self-attentive encoder. ACL, 2018(1):2676–2686.

[31] Zhang SN, Hu Y, Bian GR. 2Research on string similarity algorithm based on Levenshtein distance. In: Proc. of the IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conf. (IAEAC). 2017. 2247–2251.

[32] Benesty J, Chen J, Huang Y, *et al*. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Berlin, Heidelberg: Springer-Verlag, 2009. 1–4.

[33] Zar JH. Spearman rank correlation. Encyclopedia of Biostatistics, 2005,7.

[34] https://www.scipy.org/

附中文参考文献:

[3] 李亚超,熊德意,张民.神经机器翻译综述.计算机学报,2018,41(12):100–121.

[19] 王赞,闫明,刘爽,陈俊洁,张栋迪,吴卓,陈翔.深度神经网络测试研究综述.软件学报,2020,31(5):1255−1275. http://www.jos.org.cn/1000-9825/5951.htm [doi: 10.13328/j.cnki.jos.005951]

钟文康(1997－),男,学士,主要研究领域为软件工程,自然语言处理.

李传艺(1991－),男,博士,助理研究员,CCF 专业会员,主要研究领域为软件工程,业务过程管理,自然语言处理.

葛季栋(1978－),男,博士,副教授,CCF 高级会员,主要研究领域为软件工程,分布式计算与边缘计算,业务过程管理,自然语言处理.

唐泽(1994－),男,硕士,主要研究领域为代码摘要,API 补全.

陈翔(1980－),男,博士,副教授,CCF 高级会员,主要研究领域为软件缺陷预测,软件缺陷定位,回归测试,组合测试.

骆斌(1967－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为软件工程,人工智能.

© 中国科学院软件研究所 http://www.jos.org.cn