doi:10.3772/j.issn.2095-915x.2020.05.002

开放科学（资源服务）标识码（OSID）

基于预训练模型的机器翻译译文检测方法

田科[1,2] 张家俊[1,2]

1. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190；
2. 中国科学院大学人工智能学院 北京 100049

摘要：机器翻译译文检测任务旨在大规模文本中判别每句话是机器翻译译文还是人工翻译译文。现有的机器翻译译文检测方法大都采用统计的方法提取特征，但是基于统计的方法提取特征能力有限，严重依赖于离散的手工特征，而神经网络模型使用分布式表示，构建代价较低且能表达细粒度的句法、语义特征差别。在本文中，我们提出使用预训练语言模型和双向门控循环单元模型结合，提取机器翻译译文的语言风格、惯用词等隐层表示作为特征来检测机器翻译译文，检测结果相较之前的统计方法有很大的提升。本文尝试使用所提方法过滤混合机器翻译译文的双语语料，过滤后的语料相较原始的语料规模减小了，但是模型的性能却略有提升。

关键词：机器翻译译文；预训练语言模型；双语语料

中图分类号：G35

Machine-Translated Text Detection Method Based on Pre-trained Model

TIAN Ke[1,2] ZHANG Jiajun[1,2]

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The task of machine-translated text detection is to determine whether a sentence is translated by machine or human. Most of the existing detection models use statistical approaches for feature extraction. However, statistical methods have limited feature extraction capabilities and rely heavily on discrete manual features, while neural network models use distributed representations to implicitly express syntactic or semantic features. In this paper, we combine pre-trained language models and bi-directional gated recurrent unit model as a feature extractor. Then implicit features such as language style, idiomatic words are extracted. Experimental results show that our model significantly outperforms previous statistical methods. This paper further uses the proposed method to filter the bilingual corpus where machine-translated texts have been mixed in. The filtered corpus is

作者简介：田科（1994-），硕士，研究方向：机器翻译、机器翻译译文质量评估；张家俊（1983-），博士，研究员，研究方向：机器翻译、自然语言处理、多语言文本分析，E-mail: jjzhang@nlpr.ia.ac.cn。

doi:10.3772/j.issn.2095-915x.2020.05.002

smaller than the original one, but the performance of the model is slightly improved.
Keywords: Machine-translated text; pre-trained language models; bilingual corpus

引言

近年来，机器翻译 [1-3] 的性能取得了极大的进步，译文质量不断提高，尤其是神经机器翻译（Neural Machine Translation, NMT）的出现使得机器翻译译文质量达到新的高度。目前商用机器翻译系统已经普遍采用神经机器翻译方法，如 Google 翻译、百度翻译、搜狗翻译等，但是实践表明现有的神经机器翻译方法在长句和复杂句翻译方面还是无法与人工翻译相媲美。各大公司的免费机器翻译服务方便了使用者，同时机器翻译译文的质量参差不齐也困扰着我们，翻译质量差的译文会曲解原文的意思，造成不必要的误会。机器翻译的训练语料对于翻译模型的性能至关重要，收集高质量的双语语料一直以来都是一项繁重的工作。从网络中爬取双语语料是研究者们常用的方法，但是爬取到的语料质量参差不齐，其中有很多是商用机器翻译软件翻译的低质量译文，尤其是一些低资源语种，所以在大规模文本中过滤掉低质量的译文变得尤为重要。综上所述，自动检测机器翻译译文一直以来都是自然语言处理领域中亟待解决的一个任务，其任务是在大规模文本中检测某一句话是机器翻译译文还是人工翻译译文。

人工翻译和机器翻译在短句上表现相差无几，但是在结构复杂或者较长的句子中表现还是差距很大。一般来说，人工翻译较为灵活，翻译过程中对句子结构、语法应用、以及上下文的逻辑思想等都可以分析思考，译文不会出现语法混乱，逻辑不清等现象。有些文学性质的文章作品人工翻译可以表达出其中的韵味以及思想精髓，但是机器翻译可能会翻译的很直白，甚至出现句子结构混乱、错翻等现象。而且人工翻译可以根据译文读者的语言习惯，思维方式，风俗习惯等，把译文翻译的更符合读者的阅读思维习惯，使译文更加的地道精确。相比之下，机器翻译在翻译过程中比较中规中矩，根据预先训练好的模型进行推理翻译，灵活性不高，翻译过程中可能会出现错翻、重翻、漏翻等现象，但是机器翻译在翻译速度上相较于人工翻译有明显的优势，且可以实现一种语言到多种语言的翻译。

许多研究者在很早之前已经开始进行相关的工作，统计机器翻译时期最常用的是基于 N-gram 模型衡量句子的流畅度；还有研究者使用句法解析树的结构特性，通过计算解析树的节点是否平衡来判断一个句子是否为机器翻译译文；此外有研究者利用上下文一致性、齐夫定律、回译等方法检测机器翻译译文。之前所有的方法都基于统计的方法或机器学习方法，这些方法一般严重依赖离散的人工特征，而神经网络模型使用分布式表示，构建代价较低且能表达细粒度的句法、语义特征差别，且预训练语言模型已经学习到通用的语言表示，可以避免从头开始训练新模型。针对此任务，我们采用当前自然语言处理领域新的范式预训练语言模型作为提取特征模块，利用提取到的深层语义特征结合神经网络方法检测机器翻译译文。

本文提出的检测方法使用预训练语言模型 BERT（Bidirectional Encoder Representations from Transformer）[4] 和双向门控循环单元（Bi-directional Gated Recurrent Unit, 简称 Bi-GRU）[5] 模型提取机器翻译文本的语言风格、惯用词等隐层向量表示，并以此作为特征检测机器翻译译文。本文在不同语言、不同翻译系统翻译的语料上都进行了实验，实验结果验证了所提方法的有效性以及检测准确率和译文质量的相关性。我们的贡献如下：

使用预训练语言模型 BERT 和 Bi-GRU 等神经网络方法检测机器翻译文本，在性能上显著优于基于统计的方法。

实验分别从单语和双语两种输入方式提取译文特征，进一步证明引入源语言特征可以学习更丰富的信息表示译文的质量。

从实际出发，应用于双语语料过滤任务，筛选高质量的双语语料，过滤后的双语语料显著提升了机器翻译模型性能。最后从测试集句长、训练语料 BLEU（Bilingual Evaluation Understudy）[6] 值等方面分析了检测准确率和机器翻译文本质量的相关性。

1 相关工作

本文以提取特征的粒度对已有工作进行分类总结，分别为词级别、短句级别、句子级别以及句法结构，在下文的四个小节具体展开介绍。

1.1 词级别

Aharoni 等 [7] 设计三种特征输入 N-gram 模型中检测机器翻译译文，分别为虚词、词性以及虚词和词性混合。并且着重分析了机器翻译译文的翻译质量和检测准确率之间的相关性。Nguyen-Son 和 Echizen[8] 集成了带有噪声特征的单词 N-gram 模型，用于检测在线社交网络（Online Social Networking, OSN）消息中的机器翻译译文，这些噪声有出现在人工翻译译文中的拼写错误和口语单词等特征，还有容易出现在机器翻译译文中的未翻译的单词以及重复翻译的单词。人类编写的文本中词语使用习惯往往遵循齐夫定律（Zipfian），即使用频率最高的词语是排名第二的两倍、排名第三的三倍。Nguyen-Son 等 [9] 使用该定律检测机器翻译译文，此外，他们还提取了习语、陈词滥调、文言文和方言等特征。但是基于齐夫定律的检测方法只适用于单词分布稳定的大型文本中。

Lison 等 [10] 提出一种基于语言特征以及非语言特征的机器学习模型来检测机器翻译译文，在特定的字幕领域内分别在目标侧单语和双语数据提取特征。字幕领域的特点是除了纪录片之外，其余字幕本质上是对话式，通常在本句话中包含的词或者短语和前一句话有紧密联系。具体而言，该文发现在目标语单侧机器翻译译文通常比人工翻译译文包含更大比例的罕见词以及未知词。所以该方法在目标侧单语中使用统计语言模型检测译文中的罕见词和未知词的数量，以及对给定译文的二元组计算对数概率，然后将未知词的数量以及对数概率非常低的二元组数量作为集成特征输入到分类模型中。经过研究发现，字幕翻译还与原始数据的非语言特征有关，比如电影类型、发行年份、发行类型和电影或电视剧的原始语言等，将这些变量也作为特征输入分类模型中。在双语数据

doi:10.3772/j.issn.2095-915x.2020.05.002

中可以从源语言句子和目标翻译中共同抽取有用的特征，研究发现机器翻译译文更容易为源语言句子的字面直译，并且通常更具有接近 1 的平均比率。另一方面，人工翻译时可能会因为语言风格、文化等对译文增加语气词或者省略无用词，这就造成了字符数量比率的更大差异化。因此可以将源语言句子和目标翻译中字符数量比率作为特征，还可以通过词性标注获取源语言句子和目标翻译的依赖关系作为句法结构特征。

基于词级别的方法大都采用词频、词性等属性并结合 N-gram 模型设计离散的特征，效率高且易于实现，但是只使用词的统计特性作为特征显然是不够的，没有考虑到译文的流畅性、语法结构等特性。

1.2 短语级别

Antonova 和 Misyurev[11] 提出基于相似性的统计机器翻译（Statistical Machine Translation, SMT）系统的短语表构建算法，用于检测在网页端爬取的俄英双语数据中含有的机器翻译译文。该文认为 SMT 系统对单词的重新排序是基于规则的。因此，它们的输出与人工翻译是不同的，而这种差异可以作为特征来检测机器翻译译文，换句话说，基于统计的翻译方法会导致特定的单词和短语出现的很频繁。

Arase 和 Zhou[12] 提出一种数据驱动的方法检测网络文本中的低质量机器翻译译文，仅使用单语语料作为输入文本，以文本的流畅度、语法、非连续短语的完备性等特征训练分类器。非连续性短语特征主要关注于统计器翻译文本中的“短语沙拉”现象，即机器翻译文本中的每个短语在单独使用时语义和语法都是正确的，但是当与文本中的其他短语组合为一个完整文本时就变得不符合语法。示例插图（图 1）所示为 SMT 译文中“短语沙拉”现象。

源语言句子：这则新闻不仅在日本播出，而且在全球播出，外国人听到这个报道都很惊讶。

SMT 译文：| Of surprise | was up | foreigners flocked | overseas | as well, | they publicized not only | Japan, | saw an article from the news. |

参考答案: The news was broadcasted not only in Japan but also overseas, and it surprised foreigners who read the article.

图 1 SMT 译文中“短语沙拉”现象

从图 1 中的 SMT 译文可以看到，每个短语都是由连续的单词组成，语法正确且表达流畅，然而短语间的流畅性和语法正确性很差。如图 1 中红框所示为非连续短语“not only …, but also”的一部分，缺失了后半部分短语，缺失部分也是构成完整句子的重要组成部分。对于大多数 SMT 系统来说，这样的非连续短语是难以生成的，但是在人工翻译译文中却是容易出现的。具体地说，通过判断译文中非连续性短语的存在与否以及相应类分布来计算表示信息量的信息增益，并以此作为衡量非连续性短语的特征。同时该方法使用 N-gram 语言模

型分别对人工翻译译文和机器翻译译文进行建模，得到表示流畅度的特征。在语法特征层面，首先使用序列标注模型对人工翻译译文和机器翻译译文标注词性，再将得到的词性序列输入 N-gram 模型建模得到表示语法特征的特征。在特征空间中，使用支持向量机（Support Vector Machines, SVM）[13] 分类器来预测机器翻译译文和人工翻译译文的概率。

基于 SMT 系统的短语表相似性方法考虑了 SMT 系统和人工翻译在生成词汇和短语时的差异。基于“短语沙拉”现象的方法使用语法、流畅性、短语完备性等特征，人工特征的设计相较于词级别方法有了更大的改进，但是缺乏译文整体的评估以及语义理解。

1.3 句子级别

Nguyen-Son 等 [14] 通过计算待检测译文和回译文本的相似性作为特征检测机器翻译译文，相似性指标采用机器翻译领域常用评价指标 BLEU 值。该文所提想法基于的假设是人工翻译译文的回译文本相较于机器翻译译文的回译文本有更大的变化。示例插图（图 2）所示为一个人工翻译译文和机器翻译译文的回译差异。

人工翻译–英文：*One of the best examples of how to treat a **subject**, you **'re** not **fully aware is being** examined, **much** like a photo **of yourself** you **didn't** know **was being taken.***	机器翻译–英文：*One of the best examples of how to deal with **the subject** is that you are not completely **aware**.*
机器翻译–中文：*关于如何治疗某个主题的最好例子之一，你还没有完全清楚地被检查，就像你自己不知道的照片一样。*	机器翻译–中文：*如何处理这个问题的最好例子之一是你并不完全清楚。*
回译–英文：*One of the best examples of how to treat a **topic is that** you **have** not **been completely** examined, **just** like a photo you **don't** know.*	回译–英文：*One of the best examples of how to deal with **this problem** is that you are not completely **clear**.*

图 2　人工翻译译文和机器翻译译文的回译差异

图 2 中用粗体表示人工翻译译文和回译文本以及机器翻译译文和回译文本之间的变化，用下划线表示两者之间句法结构的变化。可以明显看到，人工翻译译文的回译文本相较于机器翻译译文的回译文本在短语和句法结构上都有更大的变化。因此，该方法基于句子层面的相似性特征采用线性分类、自适应 Boosting、基于序列最小优化的 SVM 以及随机梯度下降的 SVM 四种分类模型进行分类。

该方法基于句子层面使用当前广泛应用的 BLEU 值衡量译文和回译文本之间的差异，但是对不同系统产生的译文泛化性能较差，且特征只使用一系列 BLEU 值，特征类别较少。

1.4 句法结构

Li 等 [15] 只使用目标侧提取语言特征且这些特征独立于源语言，很多语言特征与句子的句法结构直接相关。他们在实验中发现人工翻译译文本在解析树的结构上比机器翻译译文本更加平衡。因此，他们从解析树中提取一系列基于平衡的特征训练基于线性核的 SVM 分类器。当我们在检测一个译文时，可以提取一系列有效的特征，比如句子结构、所有组成类型和名词短语的右分支节点数、所有组成类型和名词短语

doi:10.3772/j.issn.2095-915x.2020.05.002

的左分支节点数等。该方法还考虑了虚词和代词的密度，这是SMT系统通常出错的错误类型。在密度特征层面，该方法提取了整体功能字密度、限定词的密度、量词的密度、代词的密度、介词的密度、标点符号的密度、助动词的密度等特征。通常，集外词（Out of Vocabulary, OOV）的出现通常会使句子结构更加复杂。而且，像主谓不一致这样的问题也很容易被识别，该方法会融入一些基于词汇层面的特征，例如集外词的数量、根结点的孩子节点类型等。另外，我们将句子内的情感一致性作为特征进行评分，由于一个合理的句子应该在不同的词语之间具有一致的情感强度。示例插图（图3）所示为一个解析树的例子。

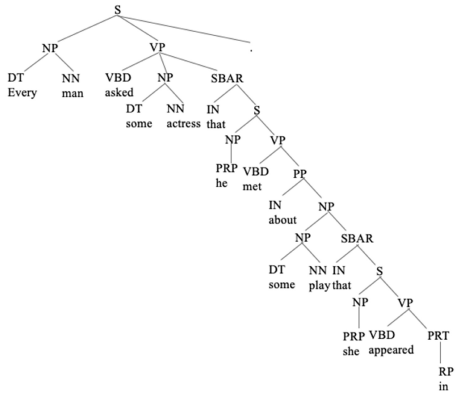

图3　解析树示例

该方法只使用目标侧语言充分利用解析树表示句法结构特征，人工设计平衡性、密度、树的分支节点比例等多种特征，但是未考虑句子流畅性等特征，且人工设计特征复杂繁琐，对不同的语言不具有普适性。

2　方法

预训练模型之前有研究者在一直探索，但真正广受关注还是在近几年。从2017年开始了预训练模型的竞赛，陆续出现了ELMo（Embeddings from Language Models）[16]，GPT（Generative Pre-training）[17]，BERT，XLM[18]等模型。本文检测机器翻译译文所采用的模型框架是BERT+Bi-GRU，在原有的参数基础上用有监督的数据进行微调。基本思路是将句子输入到BERT中，提取每句话的词向量表示，然后再通过Bi-GRU进一步强化词序信息，最后经过softmax层进行分类，输出待检测文本所对应的标签。实验主要分为两个部分，一个是单语数据的检测方法，另一个是加入源语言信息的跨语言检测方法。

2.1　BERT模型

BERT的出现为自然语言处理（Natural Language Processing，NLP）领域带来了里程碑式的改变。BERT可以视为结合了GPT和ELMo优势的新模型，其中ELMo使用两条独立训练的LSTM获取双向信息，而GPT使用新型的Transformer[19]和经典语言模型只能获取单向信息。Transformer抛弃了RNN[20]、CNN[21]等特征抽取方法，完全采用注意力机制。自注意力机制可以认为是在建立句子内部的隐藏关系，这种内部关系对序列任务有很大的帮助。

BERT的预训练过程是先从数据集抽取两个句子A与B，有50%的概率B是A的下一句，50%的概率B是一个随机选取的句子，然后将这两句话拼接输入模型中，预测句子对是否连续，这样就能学习句子之间的关系。其次是随机遮掩两个句子中15%的词，并要求模型根据上下文预测这些被遮掩的词，通过这样的辅助任务让模型学到

上下文相关的表征，并能充分利用双向的信息。

2.2 BERT + Bi-GRU

接下来两节将简要介绍所提出的方法。本文首先利用 BERT 的先验知识提取序列的隐层向量，其中隐层向量可以是 BERT 的最后一层表示，也可以是其中的某一层表示。由于 Bi-GRU 等模型天然具有时序性，适合处理序列结构数据，为了更好的表示句子中的词序信息，我们在 BERT 后面衔接了一个三层的 Bi-GRU，具体计算方法如下：

$$\vec{h}_{GRU-i} = \overrightarrow{GRU}(h^L_{BERT-i} \oplus \alpha^L_{BERT-i}) \quad (1)$$

$$\overleftarrow{h}_{GRU-i} = \overleftarrow{GRU}(h^L_{BERT-i} \oplus \alpha^L_{BERT-i}) \quad (2)$$

其中，h^L_{BERT-i}表示 BERT 第 L 层，第 i 个位置词汇的隐层向量；α^L_{BERT-i}表示 BERT 第 L 层，第 i 个位置词汇的注意力向量；⊕表示向量连接符号，将两个向量拼接在一起。$\vec{h}_{GRU-i}$和$\overleftarrow{h}_{GRU-i}$分别表示通过前向和后向 GRU 之后第 i 个位置词汇的隐层向量。

我们将两个方向的最终状态拼接后乘一个权重矩阵，最后经过 softmax 层得到标签的概率分布，计算公式如下所示：

$$p = \text{softmax}(W_0 \cdot [\vec{h}_{GRU} \oplus \overleftarrow{h}_{GRU}]) \quad (3)$$

其中 $W_0 \in R^{K*H}$ 为权重矩阵，$\vec{h}_{GRU}$和$\overleftarrow{h}_{GRU}$分别表示前向和后向两个方向的最后一个 GRU 单元的隐层向量。模型框架具体参见示例插图（图 4）。

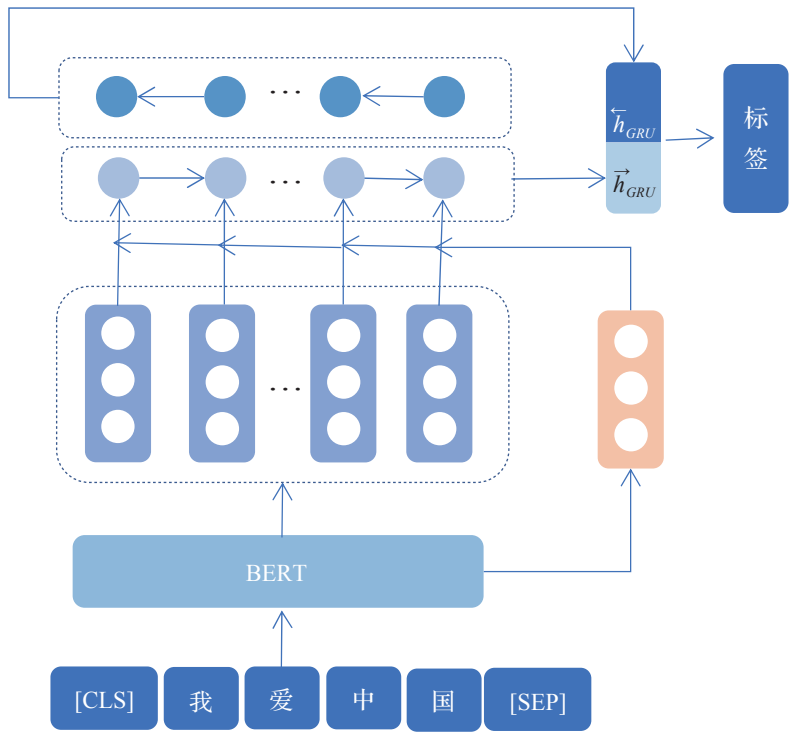

图 4 BERT + Bi-GRU 模型示意图

2.3 引入源语言特征

在过滤平行语料的任务中，我们考虑引入源语言句子特征，源语言句子的向量表征对于检测机器翻译译文有很大的帮助。平行语料即用不同的语言表达相同的语义，故我们认为源语言句子的信息可以增强目标语言

doi:10.3772/j.issn.2095-915x.2020.05.002

句子的信息量，从而进一步提高检测准确率。在引入语言特征的实验中使用 BERT 的多语言预训练模型 ---Multi-BERT，它的训练语料包含了 104 种语言的维基百科页面，并且共享了一个词汇表。首先用我们自己的数据在 Multi-BERT 的基础上继续预训练，将 BERT 原始预测是否为下一句改为预测是否是平行数据，训练完成后将平行语料的一组双语数据拼接作为输入，利用 BERT + Bi-GRU 提取双语数据的向量表征，然后经过 softmax 层计算概率分布。Multi-BERT 的输入具体参见示例插图（图 5）。

图 5 **Multi-BERT** 输入样例

3 实验结果

3.1 实验设置

在实验中分别使用了单语言和多语言的 BERT-Base 模型，其中参数为：层数 L=12，隐藏单元 H=768，多头注意头数量 N_{head}=12。在过滤平行语料的实验中，我们采用的是当前机器翻译的主流模型 Transformer，参数设置为：编码器和解码器均是层数 L=6，隐藏单元 H=512，多头注意头数量 N_{head}=8。

3.2 实验数据

为了保证实验的公平性，我们使用相同语义的人工翻译译文和机器翻译译文随机打乱顺序作为训练集，这样可以保证检测结果不依赖于语义特征。实验数据来源为联合国英中翻译语料，随机从中采样 50 万平行语料作为人工翻译的英文译文和中文译文，标签定义为 hm；调用搜狗 API 和内部翻译系统分别翻译采样的联合国语料，将翻译系统翻译后的译文标签定义为 mt。最后将人工翻译的英文（中文）译文和机器翻译的英文（中文）译文混合并随机打乱数据顺序来构建数据。为了验证所提方法在多种语言上都有效果，我们分别搜集了 50 万中日和英德数据，以同样的方法构建了相同规模的德文和日文的数据。

构建语料统计具体参见以下示例（表 1）。

表 1 构建语料统计

数据集	数量/条
训练集	1000000
验证集	2000
测试集	2000

3.3 单语数据实验结果

首先在单语数据上进行了实验，表 2 对比的是译文质量的好坏对于准确率的影响，分别使用了自己训练的翻译系统（OurNMT）和搜狗的翻译系统（SogouNMT）对源语言句子进行翻译，实验结果具体参见以下示例（表 2）。

表 2 中英文对于不同译文质量的实验对比

语言	准确率/%	召回率/%	F1值/%
中文-OurNMT	92.28	92.27	92.27
英文-OurNMT	85.37	84.98	84.17
中文-SogouNMT	87.29	86.91	87.11
英文-SogouNMT	84.15	83.83	83.99

实验结果如表 2 所示，在中文和英文两种

语言上，OurNMT 的译文检测准确率都要优于 SogouNMT 的译文检测准确率，我们认为在相同语义下，译文的检测准确率和译文的质量相关。在机器翻译领域通常使用 BLEU 值来衡量译文的质量，当译文为中文时，OurNMT 的译文 BLEU 值为 18.60，SogouNMT 的译文 BLEU 值为 36.09。当译文为英文时，SogouNMT 的译文的检测准确率相比 OurNMT 的译文的检测准确率只是微微下降，经过对比，SogouNMT 的译文和 OurNMT 的译文在 BLEU 值上相差很小。通过上述实验分析，可以看到不同质量的译文直接影响检测准确率。为了避免模型之间带来的误差，接下来我们对比在同一个翻译系统中各种语言的译文检测结果，实验结果具体参见以下示例（表 3）。

表 **3** 在同一翻译系统下各种语言的实验对比

语言	准确率/%	召回率/%	F1值/%
中文-SogouNMT	87.29	86.91	87.11
英文-SogouNMT	84.15	83.83	83.99
德文-SogouNMT	83.42	82.19	82.79
日文-SogouNMT	96.41	96.42	96.42

实验结果如表 3 所示，可以看出我们的方法在不同的语言上都有较好的表现，尤其在日文的译文检测中，F1 值达到 96.42%。通过观察结果发现，日文和中文的检测准确率都高于拉丁语系的英文和德文，我们认为这是由于不同语系的细粒度组成结构的差异性所造成。在同一语系下，日文的译文检测准确率高于中文，分析其原因是 SogouNMT 生成的中文译文质量优于日文译文。

在中文单语数据上和其他方法进行比较，由于之前方法都是在统计机器翻译生成的译文上提取特征进行实验，本文所使用的数据集中的译文为神经机器翻译生成的译文，译文质量相对较高，所以基于统计特征的方法性能有所下降。我们所提出的方法在相同数据集上达到了最好的效果。实验结果具体参见以下示例（表 4）。

表 **4** 不同方法在中文单语数据上的实验对比

方法	准确率/%	召回率/%	F1值/%
Nguyen-Son等[9]	52.43	51.94	52.35
Li等[15]	58.41	58.23	58.48
Aharoni等[7]	56.83	57.11	56.91
Nguyen-Son等[14]	80.66	80.53	80.62
Our	87.29	86.91	87.11

3.4 双语数据实验结果

本小节实验分别在中文和英文两种语言上加入源语言信息来增强译文信息量，并和单语实验结果进行对比。实验结果具体参见以下示例（表 5）。

表 **5** 引入源语言特征的实验结果对比

语言	准确率/%	召回率/%	F1值/%
中文-SogouNMT	87.29	86.91	87.11
英文+中文-SogouNMT	89.56	89.01	89.28
英文-SogouNMT	84.15	83.83	83.99
中文+英文-SogouNMT	85.29	85.06	85.18

实验结果如表 5 所示，可以发现，将源语言句子和译文拼接作为输入并把模型的输出作为特征之后，检测准确率有明显的提高，说明跨语言的信息对于检测机器翻译译文是有帮助的。分析原因是由于源语言句子和对应译文在模型内部 self-attention 进行了充分的交互，将源语言的信息融入到译文中得到更好的跨语言表示。

doi:10.3772/j.issn.2095-915x.2020.05.002

3.5 过滤平行语料实验结果

本节我们添加了一组实验来验证所提方法对于筛选高质量语料的有效性。使用引入源语言信息的方法对规模为 200 万的中英双语语料进行检测机器翻译译文，为了更接近于真实场景中机器翻译译文的分布，在双语语料中设置约 20 万机器翻译译文。经过实验检测出双语语料中有 16.9 万为机器翻译译文，将检测出的机器翻译译文从语料中删除，然后将过滤后的双语语料用于机器翻译训练。在相同的参数和模型下分别使用过滤前后的语料训练翻译模型，并对训练后的翻译模型测试性能。实验结果具体参见以下示例（表 6）。

表 6 筛选前后语料训练的翻译系统对比

NMT系统	Bleu值
原始语料训练的NMT	38.08
过滤后语料训练的NMT	40.11

从表 6 中可以发现，在相同的参数设置下，使用我们所提方法进行过滤后的双语语料训练的机器翻译模型相较于使用原始语料训练的机器翻译模型，在性能上有大约 2 个 BLEU 值的提升，进一步从侧面论证了我们所提方法对于检测机器翻译译文的有效性。

4 实验分析

4.1 分析句长对检测准确率的影响

本小节以 SogouNMT 生成的中文译文进行长度分析，首先把测试集中机器翻译译文的长度分成三个等级，短句的句长小于 15，中长句的句长大于 15 并且小于 40，长句的句长大于 40。分别统计了测试集中机器翻译译文和被错误检测的机器翻译译文的各个长度区间的数量，具体参见以下示例（表 7）。

表 7 在测试集中机器翻译译文和被错误检测机器翻译译文的不同句长数量统计

句长	机器翻译译文/%	错误检测机器翻译译文/%
短句（小于15）	207	61
中长句（大于15且小于40）	326	47
长句（大于40）	465	21

通过统计句子长度比例，可以发现短句在整个测试集中占比最低，但是在被错误检测的句子中占比却最高，长句的统计结果恰好和短句相反。易得出结论短句更容易被错误检测，分析其原因是机器翻译系统对于短句的翻译质量更高，在用词习惯、语言风格以及句法结构上都可以接近人工翻译的水平。

之后我们又进行了直观的实验，对不同句长的机器翻译译文分别进行测试，可以明显的看出长句无论在准确率、召回率还是 F1 值都显著高于短句和中等句，中等句的各项指标又高于短句，符合长度越短的译文更容易被错误检测这一结论。实验结果具体参见以下示例（表 8）。

表 8 不同句长的检测准确率

句长	准确率/%	召回率/%	F1值/%
短句（小于15）	69.27	68.73	69.01
中长句（大于15且小于40）	87.21	85.69	86.45
长句（大于40）	95.25	95.26	95.26

4.2 分析BERT模型融合Bi-GRU模型的影响

本小节主要分析 BERT 模型和 Bi-GRU 模

型的融合影响，分别在中文和英文数据集上进行了消融实验，实验结果具体参见以下示例（表 9）。

表 **9** 在中文和英文数据集上的消融实验对比

语言	方法	准确率/%	召回率/%	F1值/%
中文	BERT+-Bi-GRU	87.29	86.91	87.11
	BERT	86.83	86.15	86.59
英文	BERT+-Bi-GRU	84.15	83.83	83.99
	BERT	83.53	83.46	83.43

从实验结果可以看出，BERT+Bi-GRU 模型相比于只使用 BERT 模型，在中文和英文数据集上 F1 值均高出 0.5 个百分点。本文分析是由于 Bi-GRU 模型天然具有时序性，适合处理序列结构数据，可以更好的表示句子中的词序信息。

4.3 验证BLEU值和检测准确率的相关性

在单语数据实验时，本文已经简单分析了不同质量的机器翻译译文对于整体检测准确率的影响。接下来从具体的单个句子分析，对测试集中错误检测和正确检测的机器翻译译文进行 BLEU 值打分，具体参见以下示例（表 10）。

表 **10** 错误检测和正确检测的机器翻译译文 **BLEU** 值和数量

译文检测	句子数量/条	BLEU
错误检测译文	135	41.14
正确检测译文	863	35.6

统计结果如表 10 所示，可以发现，被错误检测的译文相比于被正确检测的译文，BLEU 值有大约 4.5 的差值。

为了更直观的观察检测准确率随 BLEU 值的大小而变化，本文添加两组不同译文质量的实验并对其进行 BLEU 值打分，通过对比来表示它们之间的相关性。如示例（表 11）所示，BLEU 值越高，译文检测准确率越低。

表 **11** 检测准确率和译文 **BLEU** 值的相关性

实验	准确率/%	Bleu值
实验1	92.28	18.6
实验2	90.11	25.1
实验3	89.04	30.26
实验4	87.24	36.09

5 结论

本研究提出了一种基于预训练模型的机器翻译译文检测方法，利用 BERT + Bi-GRU 模型来提取分布式向量表示，表示机器翻译和人工翻译的不同语言风格和用词习惯。实验结果表明，该方法在性能上显著优于基于统计的方法。为了将该方法应用于构建高质量的机器翻译训练语料，引入源语言信息来提高检测能力，过滤后的双语语料训练的翻译模型相比于原始语料训练的翻译模型，性能有一定的提高。在分析中以图表的方式展现了 BLEU 值和检测准确率的相关性，即译文 BLEU 值越高，检测准确率越低。

进一步的研究工作将从以下两个方面展开：（1）存在越罕见的语言，译文的翻译质量越差，所以将在低资源语言上进行相关实验，验证所提方法在其他语言上的效果。（2）从模型角度考量，在不降低检测准确率的前提下减小模型参数，提高检测速度。

doi:10.3772/j.issn.2095-915x.2020.05.002

参考文献

[1] 宗成庆 . 统计自然语言处理（第二版）[M]. 北京 : 清华大学出版社 , 2013.

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[3] Zhang J J, Zong G Q. Deep neural networks in machine translation: an overview[J]. IEEE Intelligent Systems, 2015, 30(5):16-25.

[4] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics. Minnesota: NAACL-HLT, 2019.

[5] Cho K, Van Merriënboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.

[6] 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. Association for Computational Linguistics, 2002:311-318.

[7] Aharoni R, Koppel M, Goldberg Y. Automatic detection of machine translated text and translation quality estimation[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014:289-295.

[8] Nguyen-Son H Q, Echizen I. Detecting computer-generated text using fluency and noise features[C]. International Conference of the Pacific Association for Computational Linguistics. Springer, Singapore, 2017: 288-300.

[9] Nguyen-Son H Q, Tieu N D T, Nguyen H H, et al. Identifying computer-generated text using statistical analysis[C]. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017:1504-1511.

[10] Lison P, Dogruöz A S. Detecting Machine-translated Subtitles in Large Parallel Corpora[C]. Proceedings of the 11th Workshop on Building and Using Comparable Corpora (LREC-2018). 2018:25-32.

[11] Antonova A, Misyurev A. Building a web-based parallel corpus and filtering out machine-translated text[C]. Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. Association for Computational Linguistics, 2011:136-144.

[12] Arase Y, Zhou M. Machine translation detection from monolingual web-text[C]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1597-1607.

[13] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data mining and knowledge discovery, 1998, 2(2):121-167.

[14] Nguyen-Son H Q, Thao T P, Hidano S, et al. Detecting Machine-Translated Text using Back Translation[J]. arXiv preprint arXiv:1910.06558, 2019.

[15] Li Y, Wang R, Zhao H. A machine learning method to distinguish machine translation from human translation[C]. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters. 2015:354-360.

[16] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[17] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL]. [2020-01-13]. https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf, 2018.

[18] Lample G, Conneau A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv:1901.07291, 2019.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Proceedings of the Annual Conference on Neural Information Processing Systems. California: NeurIPS, 2017:5998-6008.

[20] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]. Proceedings of the Third International Conference on Learning Representations. San Diego: ICLR, 2015.

[21] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]. Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017:1243-1252.