“细粒度英汉机器翻译错误分析语料库”的构建与思考

裘白莲[12]，王明文[1]，李茂西[1]，陈聪[1]，徐凡[1]
(1. 江西师范大学　计算机信息工程学院，江西　南昌　330022;
2. 华东交通大学　外国语学院，江西　南昌　330013)

摘要

机器翻译错误分析旨在找出机器译文中存在的错误，包括错误类型、错误分布等，它在机器翻译研究和应用中起着重要作用。该文将人工译后编辑与错误分析结合起来，对译后编辑操作进行错误标注，采用自动标注和人工标注相结合的方法，构建了一个细粒度英汉机器翻译错误分析语料库，其中每一个标注样本包括源语言句子、机器译文、人工参考译文、译后编辑译文、词错误率和错误类型标注；标注的错误类型包括增词、漏词、错词、词序错误、未译和命名实体翻译错误等。标注的一致性检验表明了标注的有效性；对标注语料的统计分析结果能有效地指导机器翻译系统的开发和人工译员的后编辑。

关键词： 机器翻译 ； 错误分析 ； 错误标注 ； 译后编辑

Construction of Fine-Grained Error Analysis Corpus of English-Chinese Machine Translation and Its Implications

Qiu Bailian[12] , Wang Mingwen[1] , Li Maoxi[1] , Chen Cong[1] , Xu Fan[1]
(1. School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China;
2. School of Foreign Languages, East China Jiaotong University, Nanchang, Jiangxi 330013, China)

Abstract

Machine translation error analysis, aimed at finding out problems in machine translation output, including error classes and error distribution etc., plays an important role in the research and application of machine translation. In this paper, post-editing is combined with error analysis with error labels annotated based on post-editing operations. Automatic error annotation and manual annotation are used to build a Fine-grained Error Analysis Corpus of English-Chinese Machine Translation (ErrAC), in which every annotated sample includes a source sentence, MT output, reference, post-edit, WER and error annotation. The annotated error classes include addition, omission, lexical error, word order error, untranslated word, named entity translation error etc. Annotator agreement analysis shows the effectiveness of the annotation. The statistics and analysis based on the annotated corpus can provide effective guidance for the development of machine translation system and post-editing practice.

Keywords: machine translation , error analysis , error annotation , post-editing

收稿日期：20– ; 定稿日期：20–
基金项目：国家自然科学基金(61876074，61662031，61772246); 教育部人文社科基金（16YJA740028）

第十九届中国计算语言学大会论文集，第424页-第433页，海南，中国，2020年10月30日至11月1日。

1 引言

机器翻译质量评价是机器翻译研究的重要内容。机器翻译质量评价主要有人工评价和自动评价两种方式。由于人工评价成本较高，周期较长，不易获得，目前机器翻译质量评价大多采用自动评价指标，如BLEU(Papineni et al., 2002)，METEOR(Banerjee and Lavie, 2005)和TER(Snover et al., 2006)等，这些自动评价指标依据参考译文对机器译文给出整体得分，能够反映机器翻译质量整体情况，但是无法反映机器译文具体存在哪些问题，需要在哪些方面进行改进。为获取存在问题的具体信息，就需要进行机器翻译错误分析。错误分析可以找出机器译文中具体存在的问题，有助于了解机器翻译系统的不足，找准改进的方向，还可以为机器翻译质量估计、错误预测、自动译后编辑提供参考。近十几年来，错误分析在国外机器翻译研究领域受到重视，出现很多相关的研究，例如：(Koponen, 2010)使用错误分析评价机器翻译质量，(Bojar, 2011)分析了英捷机器翻译的错误类型，(Klubička et al., 2017)通过错误分析对NMT和PBMT进行细粒度的人工评价。但在国内相关研究还较少，仅有一些针对机器译文错误进行的语言学分析。例如，(罗季美and 李梅, 2012)将机器译文错误分为词汇错译、句法错译、符号错译三大类并展开分析，(罗季美, 2014)从短语和从句层面分析了机器翻译的句法错误。这些研究仅使用独立的人工译文与机器译文做对比展开分析，而且针对的是传统的机器翻译系统如RBMT，其错误分析的结果已经不能反应当前机器翻译的水平。(孙逸群, 2019)对5篇海洋类论文摘要机辅翻译中的错误进行了剖析。其错误分析侧重实例分析和改错，而且语料规模小，不具代表性。据我们了解，目前还没有专门针对英汉机器翻译错误分析可公开获得的语料库。值得注意的是，随着神经机器翻译的发展，机器翻译质量极大提高，但是英汉翻译方向神经机器翻译质量究竟如何，还存在哪些具体问题，针对这些问题还鲜有专门的错误分析，本文尝试针对这些问题展开研究与探讨。

错误分析和译后编辑是高度相关的工作，错误分析是找出机器译文的错误，译后编辑是改正机器译文的错误。错误分析和译后编辑都可以用来评价机器翻译的质量，但以往的研究大多把错误分析和译后编辑单独使用或单独作为研究对象，较少有把两者结合起来的研究。我们将译后编辑和错误分析结合起来，先对机器译文进行译后编辑，然后以译后编辑译文(PE译文)作为参照，对机器译文进行错误标注。在此基础上，构建了一个细粒度英汉机器翻译错误分析语料库(Fine-grained **Err**or **A**nalysis **C**orpus of English-Chinese Machine Translation, 简称为ErrAC)。PE译文比参考译文更适合作为错误标注参照的原因在于，翻译本来就存在一文多译的现象，同一个源语言句子可以有多种不同的正确译文，而在机器译文的基础上进行译后编辑，力求PE译文是最接近机器译文的正确译文，其编辑距离最短。因此，以PE译文来衡量机器翻译的质量相对而言更客观，更能准确地找出机器翻译真正存在的问题。(Snover et al., 2006)研究结果表明，使用人工译后编辑译文得到的HTER值，比最接近机器译文的参考译文的TER，更能准确地衡量机器翻译的质量，而且，HTER与人工评价的相关性比BLEU与人工评价的相关性更高。下面给出了WMT19新闻机器翻译测试集上的两个实例，它们表明以人工参考译文和PE译文作为错误分析参照的区别。

例1.

源语言句子： It would be extremely ill advised to venture out into the desert on foot with the threat of tropical rainfall.
机器译文： 在 热带 降雨 的 威胁 下 ， 徒步 冒险 进入 沙漠 是 极 不 明智 的 。
PE译文： 在 热带 降雨 的 威胁 下 ， 徒步 冒险 进入 沙漠 是 极 不 明智 的 。
参考译文：由于 热带 降雨 的 威胁 ， 沙漠 冒险 活动 将 十分 危险 。
PE译文WER 0.00 参考译文WER 76.92

例2.

源语言句子： Do you think he's telling the truth to the country?
机器译文： 你 认为 他 对 国家 说 的 是 真话 吗 ？
PE译文： 你 认为 他 对 国人 说 的 是 真话 吗 ？
参考译文：你 觉得 他 对 国人 所 说 的 是 事实 吗 ？
PE译文WER 8.33 参考译文WER 35.71

第十九届中国计算语言学大会论文集，第424页-第433页，海南，中国，2020年10月30日至11月1日。

从例1可见，机器译文是正确的译文，达到了翻译的忠实、通顺的要求，但是与参考译文有很大的差别。如果按照参考译文来标注错误，那么会得出这一机器译文质量低劣的结果，其WER值(Word Error Rate，词错误率)高达76.92，这样显然无法准确、有效地衡量机器译文质量。例2中，译后编辑实际上只需要一次替换的编辑操作，即修改一处错词，就可以达到忠实、通顺的要求，PE译文WER为8.33，但是机器译文与参考译文的差别较大，WER为35.71，把机器译文修改成参考译文需要三次替换操作和一次插入操作。由此可见，使用PE译文作为参照对机器译文进行错误标注，比直接使用参考译文更客观，更能有效地反映机器译文的质量，更能准确地反映机器翻译系统的问题。

本文工作的意义体现为以下四个方面：1)获得对神经机器翻译质量更客观、更准确的评价；2)为机器翻译系统开发、译后编辑工作提供参考；3)可以为机器翻译质量估计、错误预测、自动译后编辑提供数据和参考；4)可用于错误类型与自动评价指标、译后编辑工作量之间相关性的研究。

下文结构如下：第2节介绍错误分析和译后编辑相关研究和相关语料库建设情况；第3节介绍语料来源和语料库构建过程；第4节对错误标注结果进行统计与分析；第5节总结全文。

2 相关工作

错误分析可以以人工和自动两种方式进行。(Vilar et al., 2006)建立了人工错误分析的框架，定义了错误类型，根据错误分类对机器译文进行错误标注。(Popović et al., 2006)提出基于屈折变化和句法信息的自动错误分析框架，自动获得错误的细节信息。机器翻译错误分析主要有以下几种应用。第一，用于评价某一机器翻译系统的质量(Vilar et al., 2006; Lommel et al., 2014)，或比较几种不同的机器翻译系统，通常是比较SMT和NMT等不同系统(Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Bentivogli et al., 2018)；第二，考察不同错误类型对机器翻译质量的影响(Popovic et al., 2013; Federico et al., 2014)；第三，用于译后编辑的相关研究，考察不同错误类型对译后编辑工作量不同方面的影响(Krings, 2001; Zaretskaya et al., 2016)。但是，这些研究是在机器译文上进行错误分析，或者以参考译文为参照进行错误分析，不是以PE译文为参考，这会导致错误分析与实际情况存在偏差。

随着译后编辑在翻译行业越来越普遍，逐渐出现了一些可公开获得的译后编辑语料(Potet et al., 2012)。WMT从2012年开始质量估计子任务，从2015年开始自动译后编辑子任务，这两个子任务都提供了译后编辑译文语料。部分语料还有错误标注，包括基本的编辑距离操作如替换、删除、插入和移位，或者“好”、“差”二元标签，其语言对涉及英德、英俄等。CWMT从2018年开始翻译质量估计任务，提供英汉语言对机器翻译译后编辑译文，部分语料有“好”、“差”二元标签，部分语料有每个句子的HTER值。这些语料对机器翻译质量估计、错误预测、自动译后编辑、译后编辑人员培训都非常有用。

同时，还出现了一些做了错误标注的译后编辑语料库。例如，TRACE语料库包含法英、英法译后编辑译文，其中有基本编辑距离错误类型的标注(Wisniewski et al., 2013)。(Koponen, 2012)使用英西机器翻译语料，提供译后编辑译文，对语料进行错误标注，研究错误类型与估计的译后编辑工作量、实际编辑操作之间的关系，但是其语料不能公开获得。

Terra语料库(Fishel et al., 2012)是可以公开获得的人工错误标注语料库，用于自动错误分类工具Addicter(Zeman et al., 2011)和Hjerson(Popović, 2011)的评估。这个语料库由不同研究小组独立标注，标注策略各不相同，有的小组不使用参考译文，有的小组使用参考译文。这样会导致错误标注一致性不高，因为标注策略不同，标注的结果会有较大差异。而且，这项工作中人工错误分类和自动错误分类是完全独立进行的。TARAXü语料库(Avramidis et al., 2014)能够公开获得，该语料库包含译后编辑译文和机器翻译错误标注数据，但是这两项工作是完全独立进行的，而且不是在同一个数据集上进行。PE2rr语料库(Popović and Arcan, 2016)在译后编辑译文的基础上进行错误标注，更准确地反映了机器译文的错误情况，可以公开获得，但是该语料库只包含英语、塞尔维亚语、德语、西班牙语之间的语料。这些语言均属于印欧语系，语言之间的差别相对较小，其错误分析的数据可能无法一般化。英语和汉语分属于不同的语系，差别较大，有的印欧语系语言之间机器翻译常见的错误如屈折错误在英汉语言方向上没有，而有的错误类型则可能比较突出，那么英汉机器翻译与其它相同语系语言之间机器翻译的错误情况、错误分布有没有差异，有什么差异，这就需要专门进行英汉机器翻译错误分析，而目前英

新闻类别	句子数	源语言句子词数	机器译文词数	编辑词数(%)
政治	700	15425	18622	2365(12.5)
经济	112	2473	3019	473(15.4)
社会	494	9293	10958	1532 (13.7)
体育	305	6269	8154	2144(25.7)
科教	174	3791	4670	651(13.6)
文艺	212	4783	6270	1215(18.9)
总数	1997	42034	51693	8380(16.2)

表 1: 源语言句子词数、机器译文词数及编辑词数

汉语言对机器翻译质量评价和错误分析还缺少类似的语料库。

3 语料库构建

本节介绍语料来源和语料库构建过程。语料库构建过程分为两个阶段，译后编辑和错误标注。首先由专业人士进行译后编辑，然后采用自动错误标注加人工标注的方式进行错误标注。

3.1 语料来源

我们的语料来源为WMT2019新闻机器翻译测试集英中翻译方向，该测试集包括源语言句子、机器译文和人工翻译的参考译文。我们将测试集按照新闻内容分为六类：政治、经济、社会、体育、科教和文艺。我们使用的机器译文是KSAI组(金山AI)提交的机器译文，该小组在英中机器翻译任务中人工打分排名第一。KSAI提交的机器译文是基于各种神经机器翻译模型，以Transformer作为基线系统，使用了几种数据过滤和回译作为数据清洁和数据增强的方法。最终模型是经过多模型集成、重排序、后处理的系统组合(Barrault et al., 2019)。语料库的统计信息见表1，句子数为1997个，源语言句子词数为42034，机器译文词数为51693，编辑词数为8380。编辑词数百分比是按照编辑词数与机器译文词数加漏词数量的百分比来计算的。

3.2 译后编辑

进行本次译后编辑工作的译后编辑人员为2名翻译专业教师，均精通英汉两种语言，具有丰富的翻译和译审经验。为保证译后编辑质量，在进行译后编辑之前，译后编辑人员经过多次讨论和修改，知晓此次译后编辑的目标和原则。译后编辑的目标是修改机器译文的错误，使译文达到忠实源语言句子、语句通顺的要求，质量适中即可。本次译后编辑采取轻度译后编辑的原则，即只进行最少量的必要的编辑操作以达到译文质量可接受的效果，不考虑风格、文采问题，也不考虑译后编辑人员在用词习惯、语法结构等方面的个人喜好问题。针对本次译后编辑制定五条具体指南如下：(1)力求译文意思正确、语句通顺；(2)确保没有信息增加或遗漏；(3)尽可能多地使用机器译文；(4)除非影响语义，否则不修改句子结构；(5)单纯的风格问题无需修改。

新闻类别	译后编辑操作(以WER作为编辑距离)			
	无 0	低 0-25%	中 25-50%	高 >50%
政治	29.7	51.9	12.1	6.3
经济	17.0	53.6	20.5	8.9
社会	33.8	44.1	16.0	6.1
体育	14.4	37.1	27.5	21.0
科教	23.6	57.5	16.1	2.9
文艺	19.3	47.2	20.8	12.7
总数	26.0	47.8	17.2	9.0

表 2: 译后编辑工作量等级分布

表1表明了语料库句子数量、源语言句子词数、机器译文词数，以及机器译文经过译后编辑的编辑词数。整体编辑词数百分比为16.2%，需要进行编辑修改的比例不是很大，这表明在新

闻翻译对质量要求适中的应用场景中，在领域语料比较丰富的基础上，神经机器翻译质量达到很大程度可接受的水平，机器译文在很大程度上可用。在各种新闻类别中，编辑词数百分比最大的是体育新闻，达到25.7%，是出现错误最多、需要译后编辑量最大的新闻类别。而编辑词数百分比最小的是政治新闻，为12.5%，是需要译后编辑量最小的新闻类别。可见，不同新闻类别之间机器翻译质量的差别较大，其可能的原因在于相关领域训练语料规模的大小。

我们以WER表示机器译文每个句子的编辑距离，按照编辑距离的大小将所需的译后编辑工作量(体现为所进行的实际编辑操作)分为四个等级，结果见表2。从表2中可以看出，有26%的句子已经可接受，无需任何编辑操作，47.8%的句子只需要少量编辑操作即可达到质量适中的要求，17.2%需要中等编辑工作量，只有9%的句子需要进行大量修改。ErrAC语料库中也给出了每个句子的WER值。在不同新闻类别中，体育类需要的译后编辑工作量相对较高，不需要编辑操作的比例为14.4%，低于其它所有类别，而需要进行大量译后编辑操作的比例达到21%。

3.3 错误标注

错误标注工作分两个阶段进行。首先，以PE译文为参考，使用Hjerson自动错误标注工具进行错误标注；然后，将自动错误标注的结果一一进行人工核对和修改，并细化和扩展错误类型。

进行错误标注之前先对中文语料进行预处理，采用清华THULAC分词工具进行分词。Hjerson工具以机器译文和PE译文作为输入，以词为单位进行错误标注，输出错误标注结果。Hjerson工具可以识别和标注五种类型的错误，即增词、漏词、错词、词序错误和屈折错误(动词时态/人称/情态/格/性/数)。Hjerson工具主要是针对英语、德语等印欧语系语言开发的，其中的屈折错误常出现于印欧语系语言之间的翻译中，而英汉机器翻译的目的语为汉语，汉语不是屈折语言，没有屈折错误，因此Hjerson实际上标注出来的错误有四种，即增词、漏词、错词和词序错误，在ErrAC语料库中分别以ext、miss、lex和reord表示。除漏词错误，所有其他错误均针对机器译文做标注。其中，在机器译文中出现了而在PE译文中没有出现的词标注为增词。在PE译文中出现了而在机器译文中没有出现的词标注为漏词。漏词错误需要针对PE译文做标注，因为漏词是机器译文中没有的词，无法在机器译文的标注中体现，在PE译文上做标注，才能体现漏词错误及漏词的位置。

自动标注之后进行人工标注，标注者为本文作者之一，知晓标注规则和方法。在人工标注阶段，除核对和修改自动错误标注，还对错误类型进行了细化和扩展。细化针对增词错误类型，细化的标注有两种，一是数词加量词，二是人称代词加结构助词。由于英汉语言习惯的差别，这两种增词错误是英汉翻译中经常出现的问题，在机器翻译中更为明显。英文中的冠词a或an，在机器翻译中常被译为一个、一种、一名等，而很多情况下按照汉语的习惯用法这些是应该省略的，如例3所示。数词和量词的增词分别标注为ext-num和ext-cla，其出现次数分别为81次和83次，占增词总数的4.76%和4.87%。

例3.

源语言句子： Thomas Bjorn, the European captain, knows from experience that a sizeable lead heading into the last-day singles in the Ryder Cup can easily turn into an uncomfortable ride.
机器译文： 欧洲 队长 托马斯 · 比约恩 （Thomas Bjorn） 从 经验 中 知道 ， 在 莱德杯 最后 一 天 的 单打 比赛 中 ， 一 个 相当 大 的 领先 优势 很 容易 演变 成 一 场 不 舒服 的 比赛 。
PE译文： 欧洲 队长 托马斯 · 比约恩 （Thomas Bjorn） 根据 经验 知道 ， 在 莱德杯 最后 一 天 的 单打 比赛 中 ， 大比分 的 领先 优势 也 很 容易 变 成 不利 局面 。
机器译文标注： x x x x x x x x x lex x lex x x x x x x x x x x x x x ext-num ext-cla lex lex x x x x x lex x ext-num ext-cla lex lex lex lex x
PE译文标注： x x x x x x x x x lex x x x x x x x x x x x x x x lex x x x miss x x lex x lex lex x

此外，英语中的人称代词we/he/she/they等以及其相应物主代词our/his/her/their等，在机器翻译中基本都按原本译出，但是根据汉语使用习惯，很多时候在译文中都应该省略，否则译文不自然、不通顺，如例4所示。人称代词和结构助词增词分别标注为ext-pro和ext-aux，分

别出现114次和84次，分别占增词总数的6.69%和4.93%。

人工标注阶段扩展的三种错误类型为未译、命名实体翻译错误和标点符号错误。机器译文中出现了一些未经翻译的英文单词，标注为untr。机器译文中还出现了一些命名实体翻译错误或命名实体翻译前后不一致的问题，包括人名、地名、组织结构名称等。未译的大多都是命名实体，但因为错误形式不同，所以做了区分。命名实体翻译错误标注为nen。此外，还有标点符号错误、多余或遗漏的问题，这类问题全部归类为标点符号错误，标注为punc。

例4.

源语言句子： We've transformed the look and feel of our beauty aisles to enhance the environment for our customers.
机器译文： 我们 已经 改变 了 我们 美容 通道 的 外观 和 感觉 ， 为 我们 的 客户 改善 了 环境 。
PE译文： 我们 已经 改变 了 美容 通道 的 外观 和 氛围 ， 为 客户 改善 环境 。
机器译文标注：x x x x ext-pro x x x x x lex x x ext-pro ext-aux x x ext x x
PE译文标注：x x x x x x x x x lex x x x x x x

除了细化和扩展错误类型，在人工标注阶段还进行了多标签错误标注。因为有的词存在多种错误，如错词、未译、命名实体翻译错误也可能出现在错误的位置上，即同时也是词序错误。这种情况自动错误标注工具无法标注，在人工阶段做了补充，针对叠加的词序错误标注了多错误标签，在语料库中表示为+reord。

错误标注完成之后，为检验标注质量，我们进行了标注者一致性分析。我们采用取样的方法，取数据集中前100个句子，分别由A1和A2两位标注者独立进行标注，两位标注者均经过培训，知晓标注规则和方法。错误标注不是简单的打分或排序，它涉及所标注的错误数量、错误类型和标注的位置，标注者一致性不容易计算。我们采用(Stymne and Ahrenberg, 2012)关于错误标注不同标注者一致性的计算方法，该计算方法关注所标注错误的共现情况，即

$$Agreement = \frac{2 * A^{agree}}{A1^{all} + A2^{all}} \quad (1)$$

其中上标all表示每位标注者标注的总数，上标agree表示两位标注者标注错误类型相同的数量。不同标注者一致性详见表3，整体一致性达90.6%。可见，在自动标注工具的基础上进行人工修改，不仅提高了错误标注效率，也有助于提高标注者一致性。

不同标注者一致性			
错词	88.8%	未译	100%
增词	74.9%	命名实体	100%
漏词	97.6%	标点符号	100%
词序	99.5%	总数	90.6%

表 3: 不同标注者一致性

该计算方法关注所标注错误的类型和数量，没有考虑标注错误的位置。在ErrAC语料库中，我们经过观察发现，不同标注者出现标注位置不一致的主要是词序错误，即reord的标注位置会有差异，其他错误类型的标注位置基本上差异不大。各种错误类型中，增词的标注者一致性相对较低，这是因为在英汉翻译中，词与词并不是一一对应的，词一对多、多对一的情况很常见，会造成标注者对于某个词是属于增词还是错词的标注产生差异。例如源语言句子中"holiday homes"，机器译文为"度假 之 家"，PE译文为"度假屋"，标注者A1标注为"lex lex lex"，标注者A2标注为"lex ext lex"。两者对"之"字的错误类型标注不一致，分歧的原因在于标注者A1将"度假 之 家"三个词理解为对应源语言句子"holiday homes"两个词，而标注者A2的理解是"度假"对应源语言句子"holiday"，"家"对应源语言句子"home"，那么"之"就理解为是增词。

采用同样的计算方法，我们还计算了同一标注者一致性。在标注者A1完成第一次标注之后，间隔两个月的时间，随机取数据集中100个句子再次进行标注。经过计算得出，同一标注者一致性为93.6%。

	各种错误类型数量						
新闻类别	增词	错词+词序错误	漏词	词序错误	未译+词序错误	命名实体+词序错误	标点符号
政治	500	1087 +62	326	414	56 +0	38 +1	1
经济	104	226 +16	53	56	47 +2	3 +0	7
社会	287	662 +38	237	241	98 +0	42 +0	9
体育	425	1131 +64	202	303	23 +5	124 +5	21
科教	155	262 +16	101	96	40 +2	12 +1	6
文艺	232	493 +41	165	205	88 +3	65 +1	25
总数	1703	3861 +239	1084	1315	354 +10	284 +8	87

表 4: 错误类型数量

	各种错误类型错误率(%)						
新闻类别	增词	错词+词序错误	漏词	词序错误	未译+词序错误	命名实体+词序错误	标点符号
政治	2.68	5.84 +0.33	1.75	2.22	0.30 +0.00	0.20 +0.01	0.01
经济	3.44	7.49 +0.53	1.76	1.85	1.56 +0.07	0.10 +0.00	0.23
社会	2.62	6.04 +0.35	2.16	2.20	0.89 +0.00	0.38 +0.00	0.08
体育	5.21	13.87 +0.78	2.48	3.72	0.31 +0.03	1.52 +0.06	0.26
科教	3.32	5.61 +0.34	2.16	2.06	0.86 +0.04	0.26 +0.02	0.13
文艺	3.70	7.86 +0.65	2.63	3.27	1.40 +0.05	1.04 +0.02	0.40
总数	3.29	7.47 +0.46	2.10	2.54	0.68 +0.02	0.55 +0.02	0.17

表 5: 错误率(注：错误率为错误数量与文本总词数的百分比)

4 统计与分析

我们对错误标注结果做了统计，每种错误类型的数量和错误率见表4和表5。错误率是错误数量与文本总词数的百分比，这样方便对不同的机器译文进行错误分析时相互比较。从表4可见，数量最多的错误类型是错词，即在机器翻译中选择了错误的词汇进行翻译，错词数量为3861，约占编辑词数的46%。其次是增词，数量为1703，约占编辑词数的20%。词序错误和漏词分别约占16%和13%。

错误分析对机器翻译系统开发具有很好的参考价值，其主要意义在于，有助于了解机器翻译系统存在的具体问题，了解系统的不足和短板，明确改进的方向，为机器翻译系统开发提供参考。我们对神经机器翻译译文进行错误分析，根据所发现的主要问题，对机器翻译系统开发提出建议如下。

第一，针对一词多义问题。通过错误分析可知，错词问题是神经机器翻译的主要问题。机器译文中错词问题大多是因为源语言句子中一词多义，而目前的神经机器翻译技术没有对句子进行真正的理解，无法根据领域和上下文信息来选择正确的义项，导致翻译时选词错误。建议机器翻译系统开发时，一方面通过引入外部的领域知识库或知识图谱，充分利用外部知识，另一方面通过大型单语语料库训练准确的语境词向量进行词义消岐，充分利用上下文信息，来缓解一词多义导致的错词问题。

第二，针对增词错误。在ErrAC语料库中，代词加结构助词、数词加量词这两种类型的增词占增词总数的21.25%。在机器翻译系统开发时，可以考虑对这些词类的翻译设置一定的约束，同时还需要提高训练语料的质量。如果训练语料在这些词类的翻译上处理得比较好，神经机器翻译在这方面也会有更好的表现。

第三，针对术语翻译错误。以体育类新闻为例，体育类新闻中错词的数量多达1131处，占语料库中错词总数(3861)的29.3%，其错误率为13.87%。原因在于，体育类新闻中很多词是专业术语，在译文中也需要对等地翻译成专业术语，而机器翻译往往把这些词按照常用义项译出，没有根据领域来选择合适的义项，导致翻译错误。比如，The attempt sailed high above the box，句中的“box”，机器翻译为“盒子”，而在足球术语中应为“禁区”。建议开发机器翻译系统时，引入相关领域的术语词典资源，并使系统在待译文本输入时可以识别其所属领域，即时调用相关领域术语资源，以缓解术语翻译错误的问题。

第四，针对代词引起的翻译错误问题。机器翻译中由于对代词指代对象不明，导致出现翻译错误的情况很多，有时甚至引起整个句子的意思出现偏差。代词指代不明有多种原因，比如，句中代词可指代的对象有多个，导致代词指代模糊；或者代词的指代对象距离代词很远，

跨越了单个句子。目前神经机器翻译模型大多是句子级别的，无法很好地利用篇章上下文信息解决跨越句子的指代问题。建议开发和改善以段落、篇章为输入单元的翻译模型，开发基于篇章级别的神经机器翻译系统。这样的系统还可以获取句子之间的依赖关系，更连贯地翻译整个篇章文本。

第五，针对缺乏训练语料问题。领域相关语料稀缺会直接影响翻译质量，比如，体育类新闻中命名实体翻译错误多达124处，占语料库中命名实体翻译错误总数(284)的43.7%。原因在于，体育类新闻中人名、球队名、俱乐部名称等出现的频率比其他类新闻更高。在机器译文中，这些命名实体翻译出现译错以及翻译前后不一致的情况很多。这些命名实体不能正确翻译的直接原因是相关领域的训练语料较少。针对这一问题，一方面当然是尽可能增加语料的数量，扩大训练语料的覆盖度，另一方面是提高训练语料的质量。应当避免直接从网上爬取双语语料作为训练语料，而要仔细甄别双语语料的质量，使用高质量的双语语料。获得大量高质量的双语语料对于提高神经机器翻译质量具有决定性作用。此外，针对命名实体翻译的问题，建议在机器翻译系统中加入命名实体翻译检查机制，检查并改正命名实体翻译前后不一致的情况。

从ErrAC语料库的数据中可以总结出一些经验教训供译后编辑人员参考。

第一，关注一词多义引起的错词问题。各种类型错误中，错词数量最多，达到3861次，可见一词多义仍然是机器翻译的一个障碍，目前神经机器翻译系统还无法根据领域和上下文选择正确的词义进行翻译。因此，在译后编辑过程中，需要关注一个词在不同领域、不同上下文中表达的不同意义，关注词义选择的问题，提高译后编辑的准确率和效率。

第二，善于发现和修改词序错误能有效提高译后编辑效率。词序错误占编辑词数的16%。据(Kirchhoff et al., 2012)研究发现，词序错误是机器翻译使用者最不喜欢的错误类型。其原因可能在于词序错误更难发现和修改，特别是长距离词序错误。(Popovic et al., 2014)发现，错词和词序错误所需要的认知努力最大。如果是错词叠加词序错误，需要的译后编辑认知努力更大，需要的译后编辑时间更多。因此，词序错误所需要的译后编辑工作量可能相对较大，在译后编辑过程中需要予以关注。译后编辑人员应该熟悉中英文在词序方面的差异，增强对翻译中词序问题的敏感性。

第三，在ErrAC语料库中，增词错误数量较多，但相对比较容易修改。(Popovic et al., 2014)发现，删除增词的编辑操作所需要的译后编辑认知努力和时间最少。而且，关于增词错误，还可以关注代词加结构助词、数词加量词这样的增词，在本语料库中，这几种类型的增词占增词总数的21.25%。这样有针对性地进行译后编辑，有助于提高译后编辑的速度和效率。

第四，具备全局意识，从篇章整体的角度修改错误。在机器译文中，经常出现命名实体翻译前后不一致的问题，影响篇章的连贯性，导致译文读者理解困难。虽然译后编辑人员在篇章全局的理解和把握上有优势，但有时容易忽略篇章信息，更多关注单个句子的细节。因此，在译后编辑过程中需要对该问题予以注意，修改译名不一致的问题，保证命名实体翻译前后一致，加强译文篇章的连贯性和可读性。

第五，适当关注标点符号，根据中文习惯来修改。在英汉翻译中，受英文句子结构的影响，机器译文常出现中文长句。在译后编辑过程中，需要根据中文习惯合理断句，插入标点符号，尤其是逗号。在ErrAC语料库中，插入标点符号的译后编辑操作达165次，其中大多数是插入逗号。

最后，加强对机器翻译的了解。译后编辑人员除了需要具备扎实的双语能力和翻译能力，还需要对机器翻译有较好的了解。他们需要了解机器翻译系统的不足和问题，熟悉机器译文中常出现的错误，尝试摸索总结其错误模式，并掌握有针对性的纠错方法。只有在译后编辑实践中不断积累经验，才能不断提高译后编辑的质量和效率。译后编辑人员可以充分利用机器翻译提供的便利，同时发挥人工的优势，促进人机融合翻译模式的发展。

5 总结

我们构建了一个可公开获得的细粒度英汉机器翻译错误分析语料库ErrAC，语料库中每一个标注样本包括源语言句子、机器译文、参考译文、PE译文、词错误率，以及基于PE译文所进行的错误标注。错误分析是机器翻译质量评价的重要内容，错误分析语料库可以准确、有效地评价机器翻译质量，获得关于机器译文错误类型、错误分布的数据，有助于了解目前神经机器翻译存在的具体问题，为机器翻译系统开发提供参考，明确其改进的方向。我们将译后编辑

与错误分析结合起来，对所进行的译后编辑操作进行错误标注，这比使用参考译文作为参照进行错误标注，更能准确地反应机器译文的具体问题，更符合人对机器译文错误的认知。错误分析对机器翻译系统的开发和译后编辑工作都有很好的参考作用，还可以为机器翻译质量估计、错误预测、自动译后编辑和译后编辑教学提供数据基础和参考作用。由于人工的限制，目前数据库规模还比较有限，而且只针对神经机器翻译做了错误分析，没有涉及SMT等其他系统的错误分析和相互比较。未来的工作除扩大语料库规模，涵盖更多领域和不同机器翻译系统的语料，还将基于该语料库构建初步的计算模型，用于机器翻译质量估计和自动译后编辑实验。此外，本文未涉及错误类型与自动评价指标、译后编辑工作量之间相关性的考察，未来将继续这方面的研究。

参考文献

Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Maja Popovic, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2014. The taraxü corpus of human-annotated machine translations. In *LREC*, pages 2679–2682.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based mt quality: An in-depth analysis on english–german and english–french. *Computer Speech & Language*, 49:52–70.

Ondřej Bojar. 2011. Analyzing error types in english-czech machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95(1):63–76.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Mark Fishel, Ondřej Bojar, and Maja Popovic. 2012. Terra: a collection of translation error-annotated corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 7–14.

Katrin Kirchhoff, Daniel Capurro, and Anne Turner. 2012. Evaluating user preferences in machine translation using conjoint analysis. In *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy*, pages 119–126.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Maarit Koponen. 2010. Assessing machine translation quality with error analysis. In *Electronic proceeding of the KaTu symposium on translation and interpreting studies*.

Maarit Koponen. 2012. Comparing human perceptions of post-editing effort with post-editing operations. In *Proceedings of the seventh workshop on statistical machine translation*, pages 181–190. Association for Computational Linguistics.

Hans P Krings. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Arle Lommel, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović and Mihael Arcan. 2016. Pe2rr corpus: manual error annotation of automatically pre-annotated mt post-edits. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 27–32.

Maja Popović, Hermann Ney, Adrià De Gispert, José B Mariño, Deepa Gupta, Marcello Federico, Patrik Lambert, and Rafael Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the workshop on statistical machine translation*, pages 1–6. Association for Computational Linguistics.

Maja Popovic, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgments of machine translation output. In *Proceedings of the XIV Machine Translation Summit*, pages 231–238.

Maja Popovic, Arle Lommel, Aljoscha Burchardt, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Relations between different types of post-editing operations, cognitive effort and temporal effort. In *Proceedings of the 17th annual conference of the european association for machine translation*, pages 191–198. European Association for Machine Translation Dubrovnik, Croatia.

Maja Popović. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96(1):59–67.

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier, and Hervé Blanchon. 2012. Collection of a large database of french-english smt output corrections. In *LREC*, pages 4043–4048.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, pages 223–231.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *LREC*, pages 1785–1790.

Antonio Toral and Víctor M Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

David Vilar, Jia Xu, D'Haro Luis Fernando, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *LREC*, pages 697–702.

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and analysis of a large corpus of post-edited translations: quality estimation, failure analysis and the variability of post-edition. In *Machine Translation Summit*, volume 14, pages 117–124.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Measuring post-editing time and effort for different types of machine translation errors. *New Voices in Translation Studies*, (15):63–92.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: what is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96(1):79–88.

孙逸群. 2019. 海洋类论文摘要机辅翻译错误剖析. 中国科技翻译, 32(2):31–33.

罗季美and 李梅. 2012. 机器翻译译文错误分析. 中国翻译, (5):84–89.

罗季美. 2014. 机器翻译句法错误分析. 同济大学学报: 社会科学版, 25(1):111–118.