印度语言的统计机器翻译

Sudhansu Bala Das、Divyajoti Panda、Tapas Kumar Mishra 和 Bidyut Kr. Patra

印度鲁尔克拉国家理工学院（NIT），奥迪沙邦，印度

印度瓦拉纳西印度理工学院（IIT），北方邦，印度

摘要

机器翻译（MT）系统通常旨在运用各种自然语言处理（NLP）技术，将源语言自动转换为目标语言，同时保留语境的原意。在各种 NLP 方法中，统计机器翻译（ SMT ）是一种非常流行且成功的架构，适用于低资源和高资源语言。SMT 运用概率和统计技术分析信息并进行转换。本文探讨了双语 SMT 模型的开发，这些模型用于实现英语与十五种低资源印度语言（ILs）之间的互译。首先，本文简要介绍全部 15 种语言，并说明与实验需求相关的概况。此外，作为实验的一部分，本文对用于模型构建的 Samanantar 和 OPUS 数据集，以及用于微调和测试的标准基准数据集（FLORES-200）进行了详细分析。本文提出了不同的预处理方法，以处理数据集中的噪声。为构建系统，本文采用了开源统计机器翻译工具包 Moses。本文使用“距离”重排序，旨在通过一种

短语重排序分类框架，理解语法规则和依赖语境的调整。在实验中，我们使用 BLEU、METEOR 和 RIBES 等标准指标评估翻译质量。

1 引言

技术从思想的起源走向全面的实际应用，在这一历程中不断攀登新的高峰。其中一项这样的历程正致力于消除语言障碍，以便在各个领域建立无缝的社会交流。在这方面，自然语言处理（NLP）、机器学习（ML）以及基于人工智能（AI）的语言建模（LM）等相关领域的进步，对开发完善的自动机器翻译（MT）系统贡献显著（Dorr等，2004年）。尽管人们采用了各种启发式方法，力求在译入目标语言时保留源语言的词汇和语境解释，但要实现所需的流畅性、充分性、口音和整体准确性，仍然颇具挑战（Chapelle等，2010年）。然而，随着现代NLP（基于AI）方法的出现，这一目标已成为可能；要训练出优质的翻译系统，需要高质量、大规模（即语料数量庞大）的平行语料库（源语言与目标语言的翻译对）。因此，对于在全球拥有大量数字足迹的高资源语言，只要训练充分，机器翻译系统就能取得相当不错的效果。另一方面，对于普遍缺乏认可且数字存在稀少的低资源语言，处理起来则要复杂得多。当低资源语言作为目标语言或源语言时，这种不平衡往往会导致翻译质量低下。因此，机器翻译系统需要理解这类低资源语言的句法（组合词语的规则）、语义学（词语及其组合的含义）和形态学（将词素——最小的有意义单位——构成词语的规则）（Somers，2011年）。

根据启发式范式，机器翻译模型可分为基于规则的系统（RBMT）、基于实例的系统（EBMT）、统计系统（SMT）和神经系统（NMT）（Tripathi等（2010））。每种方法都有其自身的优缺点。RBMT模型遵循一套规则，用以定义一种语言以及该语言中不同语言学单位（词、短语、句子）之间的交互关系（Jussà等（2012），Michael等（2000））。针对某一语言对的翻译所定义的这些规则集和系统，都以硬编码的方式写入

机器中。RBMT模型使用的语言信息主要是从单语（单种语言）、双语（两种语言）或多语（两种以上语言）词典中收集的源语言和目标语言信息。此外，模型还使用语法，涵盖每种语言在句法、语义和形态方面的规律。然而，由于RBMT模型十分复杂，要构建完善的模型就需要高技能、经验丰富的人力劳动，因此难以实现。此外，语言具有歧义性，解决歧义往往需要耗费更多时间和精力，在大型复杂模型中尤其如此。要让RBMT模型在日常生活中发挥作用，需要投入大量工作。因此，对比RBMT更高效的翻译系统仍然不可或缺。EBMT方法利用大量翻译实例（John（2005））。值得注意的是，EBMT模型会对双语语料库进行处理，也就是将双语语料库拆分成较小的部分，把这些部分翻译成目标语言，再重新组合成完整的译文句子。与RBMT模型类似，它们对源语言和目标语言的句法、语义及形态分析考虑不多。相比之下，SMT优于RBMT和EBMT模型，因为它不需要人工干预（Adam（2008））。这种翻译方法将基于统计的学习算法应用于大型双语语料库，从而帮助机器学习翻译。该方法还使机器能够翻译训练和测试期间未曾遇到过的句子。SMT的目标是将源语言中的输入词序列转换为目标语言。不到二十年间，它已主导学术机器翻译研究，并占据商业机器翻译领域的一部分。另一方面，神经机器翻译（NMT）通过神经网络（NN）实现（Stasimioti（2020））。与SMT不同，NMT没有独立的翻译模型、语言模型或重排序模型，而是采用一个单一的序列模型，每次确定一个词。预测依据的是源句、工作量以及此前在目标语言中生成的序列。NMT是一种基于深度学习的机器学习方法，它利用依赖词向量表示的大型神经网络。

尽管神经机器翻译在使用高资源语言进行的少数翻译实验中取得了显著成果，但研究人员仍不确定神经机器翻译是否真的能够取代统计机器翻译，以及其成功是否能推广到其他任务。最终，Michał（2016）在联合国语料库（包含15种低资源语言）上开展的实验揭示了这一事实。根据实验结果可以明显看出，

在大多数情况下，统计机器翻译的表现优于神经机器翻译，BLEU 分数也证实了这一点。许多研究人员（Lohar等人（2019）、Zhou等人（2017）、Wang等人（2017）、Castilho 等人（2017））指出，神经机器翻译在低资源语言方面存在各种不如统计机器翻译之处，例如神经机器翻译比统计机器翻译需要更多语料库和资源。与统计机器翻译相比，神经机器翻译的训练通常耗时更长。此外，研究表明，当测试数据与训练数据之间存在领域不兼容时，统计机器翻译的表现优于神经机器翻译（Xing等人（2018）、Mahata等人（2018））。长句也是统计机器翻译更擅长处理的领域。

英语和印度语言属于平行文本数据较少的语言，这促使我们开展针对印度语言的研究。本研究考察了统计机器翻译系统在低资源语言对上的有效性，其中许多语言对很少有人研究。我们实验中使用的数据集涵盖全部十五种印度语言，这是首次使用统计机器翻译对所有这些语言进行测试。因此，本工作的目标是使用统计机器翻译，为阿萨姆语（AS）、马拉雅拉姆语（ML）、孟加拉语（BN）、马拉地语（MR）、古吉拉特语（GU）、卡纳达语（KN）、印地语（HI）、奥里亚语（OR）、旁遮普语（PA）、泰卢固语（TE）、信德语（SD）、僧伽罗语（SI）、尼泊尔语（NE）、泰米尔语（TA）和乌尔都语（UR）与英语（EN）之间的双向翻译构建机器翻译系统，并检验统计机器翻译在低资源语言对上的有效性。

我们的主要目标是为低资源语言，即印度语言，开发一个可作为基线系统的机器翻译系统。以下是我们工作的主要贡献摘要：

- 据我们所知，本研究首次尝试使用 SMT，结合 Samanantar 和 OPUS数据集，研究全部十五组印度语言—英语和英语至印度语言语言对的双向机器翻译，涵盖达罗毗荼语系和印度-雅利安语族。

- 旨在从翻译角度展现印度语言的语言学特点，并通过恰当的示例讨论文字系统、书写风格和语法。

- 研究了各种数据过滤方法，以清理数据并提高翻译质量。

- 利用基于距离的重排序来检验印度语言的翻译质量。

- 通过呈现使用 BLEU、METEOR 和 RIBES 等不同自动评估指标得出的结果，可以更真实地评估翻译质量。

本文的结构安排如下。第 1.1 和 1.2 小节介绍统计机器翻译，并概述实验中使用的印度语言。第 2 节介绍一些使用印度语言进行统计机器翻译和神经机器翻译的代表性研究。第 3 节说明实验框架，包括数据集和方法的概述。第 4 节介绍一些用于机器翻译评估的主要指标。第 5 节呈现结果，随后第 6 节给出结论和未来方向。

1.1 统计机器翻译

统计机器翻译（SMT）依赖统计方法（Philipp 等，2007；Richard 等，2002；Mary 等，2011）。这是一种数据驱动技术，利用平行对齐语料库。它使用数学方程计算源语言到目标语言翻译的可能性。统计机器翻译会为概率$P(Tl\,Si)$赋值。这里，Tl是目标语言，Si是源输入。它利用贝叶斯定理确定最大概率$P(Tl|Si)$，如下所示：

$$P(Tl \mid Si) \propto P(Tl)P(Si \mid Tl) \tag{1}$$

SMT 包含三个阶段：用于计算目标语言概率的语言模型（LM）$P(Tl)$，用于估计目标语言到源语言条件概率的翻译模型（TM）$P(Si\ Tl)$，以及解码器模型（DM）。解码器模型会在可能的源语言句子中搜索概率最大的句子（Kumawat 等人（2014））。

为了计算句子的概率，LM 使用n元模型。它根据句子中某个词前面最后n个词，为该词赋予概率，并估算翻译的可能性。链式法则有助于将句子分解为条件概率的乘积。

$$P(s) = P(w_1, w_2, w_3, \ldots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3)\ldots P(w_n|w_1w_2\ldots w_{n-1}) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3)\ldots P(w_n|w_1w_2\ldots w_{n-k}) \tag{2}$$

其中，$P(s)$是在假设使用 k 元语法模型的情况下，句子 s（由单词 w_1, w_2, …、w_n 组成）的概率。它利用目标语言对的双语平行语料库

。通过计算从句子中提取的单词或短语的似然度来实现这一点。DM 是 SMT 的最终阶段，也是最关键的阶段。它通过最大化似然度，即$P(TI)P(Si \mid TI)$，帮助选择最有可能被翻译的词语。

1.2 语言偏好

印度是一个多语言国家，各邦居民使用各种地方语言。语言的多样性给彼此交流和交换信息带来了困难。此外，公共交流方面的局限也使人们在分享感受、想法、观点和事实，以及处理商务事务时感到不便。此外，互联网上有许多以英语提供的实用资源，但许多印度人因语言障碍而难以从中受益。因此，提供一种便捷的地方语言翻译方案至关重要，既能促进有效沟通，也能帮助人们利用全局资源。为实现这一目标，技术创新不断寻求高效的方法，以便利用机器实现无误翻译，因为不可能让人工译员随处可用。机器翻译需要大量资源，并需借助恰当的知识库（规则）进行训练，以提高效率，满足对无误翻译方案的需求。翻译时，理解词语的含义很重要，但仅凭词语不足以构成完整的语言。词语必须按照严格的语法规则组织成句子，而且每种语言都有自己的书写风格。本研究选取了15种常用语言（分布于印度各地）。表1介绍了实验中使用的语言及其语言特征（Ethnologue（2022））。下文将简要介绍这些语言的翻译情况。

英语（***EN***）

英语是约 45 个国家的主要语言，使用者近 11.32 亿人。英语采用罗马字母书写，字母分为大写和小写。英语使用主谓宾结构。例如（expl1）：“The poor man took food”（穷人拿了食物），而（expl2）“food took the poor man”（食物拿走了穷人）。当主语的位置

在上述句子中发生变化时，英语句子的意义和含义也会改变。

阿萨姆语（***AS***）

印度东北部阿萨姆邦有超过1500万以阿萨姆语为母语的人。这是阿萨姆邦的官方语言之一。此外，印度东北部其他邦的多个地区也有人使用阿萨姆语。它使用孟加拉语-阿萨姆语文字，并从左到右书写，也遵循主宾动语序。“Gita is eating mango”是一个英语句子，翻译成阿萨姆语后为গীতাই আম খাই আছে，其遵循主宾谓语序。গীতাই（悉多，主语）、আম（芒果，宾语）和খাই আছে（正在吃，动词）。

马拉雅拉姆语（***ML***）

喀拉拉邦以及卡纳塔克邦和泰米尔纳德邦的一些社群使用马拉雅拉姆语交流。约有3500万民众使用这种语言。它采用主宾动书写方式，并使用主格-宾格格标记顺序。马拉雅拉姆语使用马拉雅拉姆文字，从左到右书写。句子സീതയ്ക്ക് ചിത്രരചന ഇഷ്ടമാണ് 用英语表达为“Sita loves drawing”。其中，സീതയ്ക്ക് （悉多，主语）、ഇഷ്ടമാണ് （喜欢，动词）和ചിത്രര ചന（绘画，宾语）。

孟加拉语（孟加拉语）

孟加拉语是孟加拉国的主要语言，也是印度使用人数第二多的语言。超过 2.65 亿人将其作为第一或第二语言使用。孟加拉国约有 1100 万孟加拉语使用者。在印度，阿萨姆邦、特里普拉邦和西孟加拉邦等邦都使用这种语言。它属于印度-雅利安语族。在孟加拉语句子中，标准词序为主语-宾语-动词。例如，句子 েরাজি ভাত খায় 用英语表达为“Rosy eats rice”（罗茜吃米饭）。其中，েরাজি （Rosy，主语）、ভাত （米饭，宾语）和 খায় （吃，动词）。

马拉地语（马拉地语）

马拉地语属于印度梵语衍生语族，有 9500 万人在印度使用马拉地语进行交流，主要分布在印度中部和西部。马拉地语是印度使用人数第四多的语言，拥有相当庞大的母语使用者群体

。与印地语和尼泊尔语类似，马拉地语使用天城文，书写方向为从左至右。它遵循主宾谓语序。例如，句子 तो दध िपतो 在英语中意为“He drinks milk.”，其中 तो दध。是主语，而 िपतो 是动词。

िपतो 其中 तो（他，主语），（牛奶，宾语-दध

古吉拉特语（***GU***）

古吉拉特语是古吉拉特邦4500万居民使用的语言，属于印度-雅利安语族。它采用主宾谓语书写方式，并使用古吉拉特文从左到右书写。例如，在句子 તે આઇસ્ક્રીમ ખાય છે. 中，英语意思是“He is eating ice cream.”，其中 તે 是主语，આઇસ્ક્રીમ 是宾语，动词是 ખાય છે。

卡纳达语（***KN***）

卡纳塔克邦的官方语言是卡纳达语，该语言在印度其他地区也广泛使用。在印度，约有3600万人说卡纳达语并使用卡纳达文书写。尽管卡纳达语是拥有大量历史文献的达罗毗荼语，但其计算语言学资源较少，因此由于其语义和句法多样性，研究卡纳达语文献颇具挑战。卡纳达语采用主语-宾语-动词的结构。卡纳达语是一种高度黏着语。它使用从左到右的卡纳达文。例如，ರಾಮು ಶಾಲೆಗೆ ಹೋದ（主宾谓语序）对应英语“Rama went to school”。这里，ರಾಮು（罗摩，主语）、ಶಾಲೆ（学校，宾语）和ಹೋದರು（去，动词）。

印地语***(HI)***

印地语是印度的官方语言和国家语言之一。超过6.15亿人以印地语为主要语言，另有超过3.41亿人将其作为第二语言。不过，如下例所示，其句子结构为主语—宾语—动词：गीता स्कू ल जाती ह ै। 用英语来说是“Geeta goes to school”（吉塔去学校）。在这个句子中，गीता （Gita，主语）、स्कू ल （School，宾语）和जाती ह ै（Goes，动词）。印度宪法规定，联邦的官方语言应使用天城文书写的印地语。

奥里亚语***(OR)***

奥里亚语是印度东部奥迪沙邦的主要语言。奥里亚语属于东印度-雅利安语支。

其标准语序为主语-宾语-动词（SOV），书写时使用

奥里亚文字从左到右书写。

旁遮普语（***PA***）

旁遮普语采用主语-宾语-动词语序，在印度和巴基斯坦使用，英国、阿拉伯联合酋长国、马来西亚、美国、南非和加拿大也有少数小型群体使用。它使用两种文字系统书写：西部的波斯-阿拉伯字母沙穆基文和东部的古木基文字。古木基文从左到右书写，而沙穆基文则以相反方向书写。我们̇ 印度 人̇ 在英语中意为“我们是印度人”，其中我们̇（主语），印度 （是，动词），人̇ （宾语）。

泰卢固语（***TE***）

泰卢固语是印度南部两个邦——安得拉邦和特伦甘纳邦——的官方语言。美国、加拿大和英国的泰卢固语移民群体也使用泰卢固语。泰卢固语的文本结构采用主语-宾语-动词语序，并从左到右书写。她 我ᳪ 打ᳪ了 在英语中意为“她打了我”，其中她（主语），我ᳪ（宾语），而打ᳪ了（动词）。

信德语（***SD***）

信德语在巴基斯坦有 2500 万人使用，在印度有 500 万人使用。在巴基斯坦，信德语使用改良版波斯-阿拉伯字母书写（从右向左）；而在印度，则使用多种文字系统书写，例如天城文、胡达巴迪文和古木基文（从左向右）。其语序为主语-宾语-动词。例如，句子“帕尔塔喜欢书”写作 اپٿر ڪي ڪنبات اسن رپيا آيه ，其中 اپراٿ （帕尔塔，主语）、ڪي ڪنبات （书，宾语）和 اسن رپيا آيه （喜欢，动词）。

尼泊尔语（***NE***）

它是尼泊尔的官方语言和通用语，也有印度的一些社群使用。尼泊尔语使用从左向右书写的天城文文字系统。它采用主宾谓语序。例如，“悉多吃了苹果”翻译成尼泊尔语后

变成 सीताले स्याउ खाइन्。这里，सीताले（悉多，主语）、स्याउ（苹果，宾语）和खाइन्（吃了，动词）。

僧伽罗语（***SI***）

大多数斯里兰卡人以僧伽罗语为母语。僧伽罗语属于印度-雅利安语族，在语法结构、形态变化以及主宾谓语序（SOV）方面与英语不同。僧伽罗语使用从右向左书写的僧伽罗文字。例如，“帕万写一封信”用僧伽罗语写作 පවන් ලිපියක් ලියයි，其中 පවන්（帕万，主语）、ලිපියක්（一封信，宾语）和 ලියයි（写，动词）。

泰米尔语（***TA***）

泰米尔语主要通行于印度南部的泰米尔纳德邦，也通行于泰米尔侨民众多的国家和地区，例如斯里兰卡、马来西亚和新加坡。泰米尔纳德邦内部的南部、西部和北部口语存在音系差异。泰米尔语属于达罗毗荼语系南部分支，拥有可追溯至两千多年前的丰富文学传统。在印度和斯里兰卡使用的泰米尔语是两种不同的方言。泰米尔语采用主宾谓语序。例如，“我喜欢画作”用泰米尔语写作 எனக்கு ஓவியங் கள் பிடிக்கும் ，其中“我”எனக்கு（我，主语）、விலங் குகள் （画作，宾语）和பிடிக்கும் （喜欢，动词）。

乌尔都语***(***乌尔都语***)***

它是巴基斯坦的官方语言，在印度也广泛使用。在巴基斯坦和印度，有超过1.7亿人使用乌尔都语，在英国、美国和阿拉伯联合酋长国的一些社群中也有人使用。乌尔都语的文字系统是经过修改和修订的波斯-阿拉伯字母。乌尔都语的书写结构是主语-宾语-动词。例如，“她读一本书”用乌尔都语表示为 وہ ایک کتاب پڑھتی ہے ，其中 وہ（她，主语）、ایک کتاب（书，宾语）和 پڑھتا ہے（读，动词）。

表1：机器翻译实验所用语言的语言特征

语言	文字系统	Word 语序	语系	数 使用者人数（百万）	书写 方向
阿萨姆语 (AS)	孟加拉语	主宾谓	型印欧语	15	从左到右
马拉雅拉姆语（ML）	马拉雅拉姆语	SOV	达罗毗荼语系	38	从左到右
孟加拉语（BN）	孟加拉语	SOV	印欧语系	265	从左到右
马拉地语（MR）	天城文	主宾谓	型印欧语	95	从左到右
古吉拉特语（GU）	古吉拉特语	SOV	印欧语系	60	从左到右
卡纳达语（KN）	卡纳达语	SOV	达罗毗荼语系	36	从左到右
印地语（HI）	天城文	SOV	印欧语系	615	从左到右
奥里亚语（OR）	奥里亚语	主宾谓	型印欧语	38	从左到右
旁遮普语 (PA)	波斯-阿拉伯字母, 古木基文	SOV	印欧语系	125	从右向左 从左到右
泰卢固语（TE）	泰卢固语	SOV	达罗毗荼语系	93	从左到右
信德语（SD）	天城文 波斯-阿拉伯文	主宾谓	型印欧语	25	从左到右 从右向左
僧伽罗语 (SI)	僧伽罗语	SOV	印欧语系	17	从左到右
尼泊尔语（NE）	天城文	主宾谓	型印欧语	24	从左到右
泰米尔语(TA)	泰米尔语	SOV	达罗毗荼语系	81	从左到右
乌尔都语 (UR)	Urdu	主宾谓	型印欧语	170	从右向左
英语（EN）	罗马字母 形式	主谓宾	语序 印欧语系	1,132	从左到右

2 相关工作

本节讨论了几项使用印度语言进行统计机器翻译的研究。

Dasgupta 等（2004年）讨论了一种将英语（EN）翻译为孟加拉语（BN）的机器翻译技术，该技术利用英语句子的句法，同时尽量缩短翻译时间。在生成目标句子的过程中，他们使用词典来识别宾语和主语，以及人称、数等其他实体。

（Ananthakrishnan 等人（2009年））利用机器翻译中的形态和句法预处理，创建了英译印地语（英语-印地语）统计机器翻译系统

模型。在他们的研究中，为进行形态处理，印地语中的后缀会先被切分，然后再按照印地语句法对英语源句进行重排。

2010年，Zbib 等人（2010年）在美国麻省理工学院开展了相关研究，利用统计机器翻译中的语法结构和阿拉伯语—英语新闻语料库，以获得更好的翻译结果。Kumar 等人（2015年）采用《圣经》语料库中的 20,000 个句子，开展了基于统计机器翻译的卡纳达语到英语的机器翻译研究，取得了显著成果，BLEU 分数达到 14.5，这一结果也得到了 Papineni 等人（2002年）的支持。Kaur 等人（2011年）提出了一种基于统计机器翻译的英语（EN）到旁遮普语（PA）的翻译模型，使用其自建的双语语料库，其中包含两种语言的 3844 个名称；BLEU 和词准确率分别为 0.4123（范围为 0–1）和 50.22%。

Nalluri 等人（2011年）利用约翰斯·霍普金斯大学开源架构（Li 等人（2009年））创建了“enTel”，这是一个基于统计机器翻译的英语到泰卢固语（TE）机器翻译系统。为训练该翻译系统，他们使用了少数民族语言工程计划（EMILLE）中的泰卢固语平行数据集。
2014年，Randil 等人（2014）创建了一个用于僧伽罗语（SI）-泰米尔语（TA）机器翻译系统的统计机器翻译框架。他们的研究展示了基于统计机器翻译的语言对间翻译结果，包括泰米尔语-僧伽罗语和僧伽罗语-泰米尔语。使用统计机器翻译模型进行的实验表明，僧伽罗语-泰米尔语语言对的结果比泰米尔语-僧伽罗语语言对更为显著。对于亲缘关系密切的语言，统计机器翻译表现出显著的效果。

2017年，Khan 等人（2017）对印度语言—英语机器翻译模型开展的调查揭示了统计机器翻译在 8 种语言上的重要性，这些语言包括印地语（HI）、孟加拉语（BN）、古吉拉特语（GU）、乌尔都语（UR）、泰卢固语（TE）、旁遮普语（PA）、泰米尔语（TA）和马拉雅拉姆语（ML）。他们使用了 EMILLE 语料库（Nalluri 等人（2011）），并优先采用 Moses 统计机器翻译模型构建翻译模型，将未登录词（OOV）音译为英语。在他们的研究中，以 BLEU、NIST 和未知词计数为指标进行评估，结果显示统计机器翻译的整体表现令人满意（旁遮普语-英语和乌尔都语-英语模型表现最佳，印地语-英语和古吉拉特语-英语模型表现最差）。Islam 等人（2010）提出了一个英语-孟加拉语统计机器翻译系统。他们提出了一个音译模块来处理未登录词（OOV）。为解决英语和孟加拉语之间系统性的语法差异，他们还加入了一个介词处理模块。他们使用 BLEU、NIST 和 TER 分数来检验其系统的有效性——

tem.
如今，神经机器翻译因其推动机器翻译发展并显著提升质量而广受认可。因此，许多研究人员对低资源语言和高资源语言的两种技术进行了比较。

Antonio 等人（2017）使用统计机器翻译和神经机器翻译系统，针对九种语言方向，从多个维度进行了全面评估。在他们的实验中，对于长句，统计机器翻译系统的表现优于神经机器翻译。近期，Castilho 等人（2017）采用自动评价指标和专家译者，对神经机器翻译和统计机器翻译进行了全面的定量与定性比较。根据他们的实验，统计机器翻译表现更好。

Acharya 等人（2018）利用包含 6535 个句子的尼泊尔国家语料库（NNC），对尼泊尔语（NE）的神经机器翻译和统计机器翻译进行了比较。研究人员通过实验表明，在小规模语料库上，统计机器翻译模型的表现优于基于神经机器翻译的系统，BLEU 分数达到 5.27。
2021 年，Singh 等人（2003）使用结合注意力机制的长短期记忆网络（LSTMs），基于 WAT 语料库开展实验，取得了 15.7 的 BLEU 分数，高于 14.5 的基线 BLEU 分数。

Abujar 等人（2021）利用 AmaderCAT 语料库，采用序列到序列（序列到序列）架构（一种特殊类型的循环神经网络）开发了孟加拉语-英语机器翻译模型，并构建了翻译系统，取得了 22.3 的 BLEU 分数。

2021 年，Akshai 等人（2021）展示了使用 SMT 和 NMT 进行英语和印地语到泰米尔语翻译的成果。他们的实验揭示了 NMT 的缺点，例如在处理领域术语和未登录词（OOV）短语时会出现大量错误。NMT 经常会为多义词选择不准确的词汇，在翻译词语和领域术语时偶尔还会出现重排序错误。NMT 模型生成的译文大多包含对先前转录词语的重复、奇怪的翻译，以及许多与原句毫无关联的意外句子。

3 实验框架

3.1 数据集

模型构建使用了 Samanantar 和 OPUS 数据集，测试则采用标准基准数据集 Flores 200。Samanantar 是印度语言（ILs）中规模最大的语料库集合（Gowtham 等人（2022））。该集合包含超过 4500 万个英语与 11 种印度语言的句对。实验中，Samanantar语料库用于阿萨姆语（AS）、马拉雅拉姆语（ML）、孟加拉语（BN）、马拉地语（MR）、古吉拉特语（GU）、卡纳达语（KN）、印地语（HI）、奥里亚语（OR）、旁遮普语（PA）、泰卢固语（TE）和泰米尔语（TA）。OPUS 是一个可免费获取平行语料库的大型资源。该语料库涵盖多个领域的数据，涉及 90 多种语言（Tiedemann（2012））。OPUS语料库用于僧伽罗语（SI）、信德语（SD）、乌尔都语（UR）和尼泊尔语（NE）。表 2 列出了实验所用数据集的统计信息。
FLORES-200（Marta 等人（2022））数据集是一个包含 200 种语言的多语言平行数据集，这些语言被用作人工翻译基准。该数据集由两个语料库组成，分别标记为“开发集”（997 行）和“开发测试集”（1013 行）。“开发集”数据集用于微调，“开发测试集”数据集用于测试。

3.2 方法

我们提出的流程包括以下主要步骤：

1. **设置统计机器翻译系统** 我们使用 Moses统计机器翻译工具包来构建统计机器翻译系统。它采用 C++ 和 Perl 编写。目前，这是可用的最佳统计机器翻译工具之一。首先，在 Ubuntu 中安装 Moses、GIZA++ （Och（2003））、CMPH（用于二值化）和 SRILM。对于训练、微调和测试流程，系统除了需要语言对的平行语料库外，还需要根据开发者的选择遵循可配置的阶段。

2. **数据预处理** 高质量语料库在任何机器翻译任务中都发挥着重要作用。从各种来源获取语料库时，其数据质量（对机器翻译系统的有效性至关重要）始终无法得到保证。因此，在使用数据训练统计机器翻译模型之前，去除不必要的噪声是一项重要工作。我们采用以下流程对数据进行预处理和清理：

表 2：平行语料统计

英语到印度语	平行语料库（句子）
阿萨姆语（AS）	0.14M
马拉雅拉姆语（ML）	5.85M
孟加拉语(孟加拉语)	8.52M
马拉地语(马拉地语)	3.32M
古吉拉特语(古吉拉特语)	3.05M
卡纳达语(KN)	4.07M
印地语(HI)	8.56M
奥里亚语(OR)	1.00M
旁遮普语(PA)	2.42M
泰卢固语(TE)	4.82M
信德语(SD)	195万
僧伽罗语（僧伽罗语）	868万
尼泊尔语（尼泊尔语）	335万
泰米尔语（泰米尔语）	5.16M
乌尔都语(乌尔都语)	8.95M

- **数据清理与格式化** 数据清理的目标是查找并修正语料库中的错误数据，或将其删除。在此过程中，会移除既不用于印度语言也不用于英语的字符。扩展 Unicode 中的某些标点符号会转换为对应的标准字符。印度语言语料库中的数会从英语形式转换为印度语言文字。此外，还会从语料库中移除语言对标准字母表之外的字符、多余空格和不可打印字符。本研究采用的预处理技术总结如下：

  – 删除不可打印字符 – 删除语言对之外的字符 – 删除多余空格 – 去除带重音字符的重音 – 将两个语料库中的非标准 Unicode 标点符号替换为其标准对应形式 – 将不常见标点符号替换为更常见的标点符号 – 将数字转换为统一数字系统和文字系统

3. **分词：** 这是根据给定的字符序列和指定的文档单位，将字符序列划分为更小单位（称为词元）的过程。在本例中，单词、标点符号和数字都作为词元。语料库使用经过修改的 Moses 分词器进行分词（Koehn 等，2007）。语料库中多余的标点符号（引号、撇号和逗号）也会被删除。

4. **训练真值大小写模型：** 这是为大小写使用不正确或未区分大小写的文本添加大小写信息的过程（Lita 等，2003）。使用真值大小写可以减少数据稀疏性。在训练数据集上训练真值大小写转换模型（该模型会将句子开头的词改为最常见的大小写形式）。此处使用 Moses 真值大小写转换。

5. **训练语言模型和翻译模型**：在 MOSES 中，训练过程利用词语和片段的出现情况，建立目标语言与源语言之间的联系。语言模型和翻译模型在训练数据集上进行训练并二值化。词对齐采用 GIZA++ grow-diag-final-and 对齐方法，该方法从两种对齐的交集开始，然后添加额外的对齐点。

grow-diag-final-and 模型从源到目标和从目标到源的对齐结果交集开始，然后通过两个步骤添加额外的对齐点（Och，2003）：

grow-diag：对于测得的对齐中的每个相邻点，如果源词或目标词尚未对齐，但存在于对齐的联邦中，则将该相邻点纳入对齐。

final：如果联邦中存在未对齐的短语对，则将该点添加到对齐中。

- **词对齐模型：**单词预处理完成后，下一步是进行词对齐。本研究采用 GIZA++ （Och（2003））集成的 IBM 模型来完成词对齐过程。GIZA++ 模型评估每个句子中源语言和目标语言各个单词之间词到词对齐的可能性。为了生成高质量的词对齐，

对齐是通过一系列连续估计生成的。处理句子数量较多的语料库需要数小时。对齐方法的结果建立了目标词与源词之间的对应关系。

- **重排序** 这是调整一个自然语言句子的词序，使其更接近另一种自然语言句子词序的过程。对于句法结构不同的语言而言，这是转录中的一项关键任务。Moses 系统会在训练过程中为每个短语学习不同的重排序可能性。该模型不采用默认重排序，而是使用距离重排序模型（Kumawat 等人（2014年））。

  – **基于距离的重排序：** 目标输出短语的重排序由相对扭曲概率分布re (S_t, E_{t-1})表示。这里，S_t 指被译为第t 1 个目标短语的源短语的起始位置。重排序距离（$S_t - E_{t-1}$）按如下方式计算：当源词未按顺序选取时，重排序距离就是被跳过的词数（向前或向后均可）。如果两个短语按顺序翻译，则$t = E_{t-1} + 1$；也就是说，该短语的第一个词紧接在前一个短语的最后一个词之后。在这种情况下，重排序代价为$re(0)$。基于距离的模型对重排序距离赋予线性代价，这意味着短语移动的距离越远，代价越高。

6. **微调**：这是为翻译模型确定最佳配置文件设置的过程，以便将其用于特定目的。该过程使用翻译模型翻译调优集中的所有 15 种印度语言源语言短语。然后，将模型输出与一组参考译文（人工翻译）进行比较，并调整设置以提高翻译质量。此过程会重复多次。调优过程每次迭代都会重复这些步骤，直至翻译质量达到最佳。该模型是在预处理后的 Flores-200 开发集上进行微调的。

7. **翻译**：使用最终模型将预处理后的

Flores-200 开发测试集从源语言翻译为目标语言。

8. **后处理与去标记化**：移除多余的标点符号（引号、撇号和逗号），并使用 Moses 去标记化器对翻译文件进行去标记化。

9. **评估**：我们的实验采用的评估指标包括 METEOR（Banerjee 等人（2005年））、RIBES（Wołk 等人（2016年））和 BLEU（Papineni 等人（2002年））。

4 机器翻译评估的 4 项关键指标

机器翻译评估是任何机器翻译系统中最关键的阶段。分析机器翻译任务时，既可采用自动方法，也可采用人工方法。系统输出的有效性既可以通过人工评估直接评定，也可以通过阅读案例、其他下游活动，甚至估算纠正输出所需的工作量来间接评定。人工评估能取得更好的结果，其方式包括基于任务的评估、流畅性和充分性评分、针对翻译任务的人工投票、译后编辑措施等。然而，人工评估面临的主要挑战是耗时、缺乏可重复性以及成本高昂。为了评估机器翻译输出的有效性，人们采用了多种自动评估方法，例如显式排序翻译评估指标（METEOR）、双语评估替代品（BLEU）、Levenshtein、基于排序的直观双语评估分数（RIBES）、词错误率（WER）和 NIST。自动评估指标还有一些直观优势，例如能够为同义词或释义计分。下面将讨论本文采用的几种评估指标。

1. **双语评估替补（BLEU）：** 机器翻译（MT）评估中最广泛使用的方法称为 BLEU。该方法于 2002 年首次提出（Papineni 等人（2002 年）），用于将假设译文与一个或多个参考译文进行比较。当假设译文与参考译文有许多字符串相匹配时，机器翻译评估就会给它更高的分数。BLEU 系统为译文赋予 0 到 1 之间的分数。不过，它

通常以百分比表示。译文越接近 1，就越符合参考译文。在两个数据集中，译文的匹配都是按照相同词序逐词进行的。使用 SacreBLEU 计算基线模型的 BLEU 分数。

2. **基于排序的直观双语评估分数（RIBES）：** 该指标通过在一元匹配之前纳入秩相关系数来计算，因此无需使用更高阶的 n 元匹配。此指标关注词序。为比较统计机器翻译和参考译文，它采用基于词序的肯德尔 τ 系数（τ）来表示秩差异（Wołk 等人（2016））。为确保系数为正值，会按如下方式对其进行归一化：

$$\text{Normalized Kendall's } \tau \text{ (NKT)} = \frac{\tau + 1}{2} \tag{5}$$

该系数可与一元精确率p_1 和简短惩罚BP 结合，并进行调整，以防止高估统计机器翻译与参考译文中仅相关词语之间的相关性。

$$\text{RIBES} = \text{NKT}.(p_1^{\alpha}).(\text{BP}^{\beta}) \tag{6}$$

其中，α 和 β 是介于 0 和 1 之间的参数。

3. **显式排序翻译评估指标（METEOR）：** METEOR 根据译文与给定参考译文之间明确的逐词相似度对翻译进行评分（Banerjee 等人（2005年））。它专门用于生成与人工评估翻译质量高度相关的句子级分数。METEOR 将召回率与精确率结合使用并突出召回率；许多研究已证实，这一特点对于与人工判断高度相关至关重要。它还旨在通过采用灵活的词语匹配，并考虑同义词和形态变化，解决参考译文不够精确的问题。要获得 1 分，机器生成的输出中的词语必须出现在参考译文中，且参考译文中的每个词语也必须出现在机器生成的输出中。

5 结果与讨论

本研究采用的评估指标包括 METEOR（Banerjee 等人（2005年））、RIBES（Wołk 等人（2016年））和 BLEU（Papineni 等人（2002年））。本研究使用的所有评估指标都是衡量机器翻译文本质量的重要指标。

表 3 展示了在未进行微调的情况下，使用统计机器翻译将全部 15 种印度语言翻译成英语及反向翻译的结果。使用 Flores-200 开发集进行微调的统计机器翻译评估指标见表 4。RIBES 和 METEOR 的取值范围为 0–1。使用统计机器翻译进行英语至印度语言和印度语言至英语翻译时，BLEU 分数分别介于 0.46 至 13.09 和 0.49 至 15.41 之间。英语至印度语言和印度语言至英语的 RIBES 分数分别介于 0.04 至 0.63 和 0.14 至 0.61 之间。英语至印度语言的 METEOR 分数介于 0.01 至 0.28，印度语言至英语的分数介于 0.02 至 0.28。采用距离重排序技术的统计机器翻译模型在孟加拉语、旁遮普语、乌尔都语、印地语和古吉拉特语上的 BLEU 分数高于其他语言。未进行微调时，僧伽罗语在两个翻译方向上的三项指标均为所有语言中表现最差；而进行微调后，英语-僧伽罗语和泰米尔语-英语模型在各自方向上均逊于其他所有英语至印度语言和印度语言至英语模型，且

表 3：未进行微调的统计机器翻译评估指标结果

语言	语言对	BLEU	RIBES	METEOR
AS	英语-阿萨姆语	1.90	0.50	0.09
	阿萨姆语-英语	3.21	0.46	0.11
ML	英语-马拉雅拉姆语	3.79	0.27	0.08
	马拉雅拉姆语-英语	4.59	0.43	0.12
BN	英语-孟加拉语	6.41	0.62	0.17
	孟加拉语-英语	3.06	0.45	012
MR	英语-马拉地语	3.17	0.43	0.09
	马拉地语-英语	3.62	0.43	0.09
GU	英语-古吉拉特语	7.62	0.56	0.16
	古吉拉特语-英语	10.14	0.59	0.21
KN	英语-卡纳达语	5.06	0.40	0.11
	卡纳达语-英语	7.17	0.51	0.16
HI	英语-印地语	13.09	0.63	0.28
	印地语-英语	15.41	0.64	0.28
OR	英语-奥里亚语	3.92	0.59	0.14
	奥里亚语-英语	6.41	0.52	0.17
PA	英语-旁遮普语	7.22	0.63	0.18
	旁遮普语-英语	11.7	0.61	0.24
TE	英语-泰卢固语	8.16	0.42	0.12
	泰卢固语-英语	5.77	0.52	0.18
SD	英语-信德语	1.29	0.39	0.08
	信德语-英语	2.48	0.35	0.09
SI	英语-僧伽罗语	0.93	0.05	0.02
	僧伽罗语-英语	0.49	0.14	0.05
NE	英语-尼泊尔语	6.00	0.58	0.16
	尼泊尔语-英语	8.29	0.53	0.19
TA	英语-泰米尔语	2.78	0.16	0.05
	泰米尔语-英语	2.64	0.31	0.07
UR	英语-乌尔都语	9.43	0.62	0.24
	乌尔都语-英语	11.35	0.61	0.23

表 4：微调后的统计机器翻译评估指标结果

语言	语言对	BLEU	RIBES	METEOR
AS	英语-阿萨姆语	2.17	0.50	0.08
	阿萨姆语-英语	3.21	0.42	0.10
ML	英语-马拉雅拉姆语	2.05	0.23	0.06
	马拉雅拉姆语-英语	1.84	0.27	0.06
BN	英语-孟加拉语	8.26	0.63	0.19
	孟加拉语-英语	12.16	0.60	0.23
MR	英语-马拉地语	2.43	0.39	0.08
	马拉地语-英语	2.49	0.36	0.07
GU	英语-古吉拉特语	5.82	0.52	0.14
	古吉拉特语-英语	3.56	0.45	0.01
KN	英语-卡纳达语	3.35	0.09	0.14
	卡纳达语-英语	3.67	0.41	0.10
HI	英语-印地语	8.64	0.57	0.22
	印地语-英语	5.38	0.49	0.14
OR	英语-奥里亚语	5.25	0.58	0.15
	奥里亚语-英语	2.22	0.39	0.11
PA	英语-旁遮普语	5.71	0.60	0.15
	旁遮普语-英语	7.75	0.55	0.19
TE	英语-泰卢固语	4.4	0.38	0.10
	泰卢固语-英语	3.34	0.44	0.12
SD	英语-信德语	1.59	0.41	0.09
	信德语-英语	2.53	0.38	0.09
SI	英语-僧伽罗语	0.46	0.04	0.01
	僧伽罗语-英语	3.11	0.37	0.11
NE	英语-尼泊尔语	4.00	0.55	0.14
	尼泊尔语-英语	5.25	0.49	0.13
TA	英语-泰米尔语	1.86	0.16	0.05
	泰米尔语-英语	1.03	0.08	0.02
UR	英语-乌尔都语	6.34	0.56	0.19
	乌尔都语-英语	7.07	0.54	0.18

三项指标。与其他语言相比，HI 和 BN 的数据集质量较高、规模较大且噪声较少。因此，在不进行微调的情况下，HI 在两个方向的三项指标上都优于所有其他语言；而在进行微调后，BN 在两个方向的 BLEU 和 RIBES 指标上均优于所有其他语言。此外，UR 和 PA 的 RIBES 指标也优于其他语言。PA 的 RIBES 分数为 0.63（英语-旁遮普语）和 0.61（旁遮普语-英语）；UR 的 RIBES 分数为 0.62（英语-乌尔都语）和 0.61（乌尔都语-英语）。

尽管 SI 的语料库规模不小，但与其他语言相比，其中的翻译并不可靠。例如，英语句子“希伯来书 11:32-34；士师记 16:18-21，28-30 耶和华的灵在特殊情况下以独特的方式作用于参孙”在语料库中被译成僧伽罗语“ 11:32-34；ගී.”，其中只翻译了“希伯来书 11:32-34；”。因此，与其他语言相比，SI 的表现不佳。同样，在英语-泰米尔语语料库中，句子“He’s my boss”被译为“அவர் எனது ேமலாளர் மட்டும் தான் .”，实际意思却是“他只是我的经理”。从这个例子可以看出，英语-泰米尔语语料库也存在歧义。此外，尽管印度语言-英语和英语-印度语言系统使用相同的语料库进行训练，但 BLEU 分数却存在显著差异。这是由于印度语言具有显著的形态多样性，而且从英语翻译成印度语言相对困难。研究发现，SI 的行数很多（8.68 M），但其表现不如 PA（2.42 M）和 GU（3.05 M）等语言。还发现，斜率非常陡的语言往往得分较低。例如，英语-泰米尔语和英语-马拉雅拉姆语中，句长少于 4 个词元的句子占 60%，其得分不太理想，如图1所示。相较之下，HI 和 BN 等得分较高的语言，其斜率较为平缓。因此，句子长度也是一个影响因素。英语-信德语是个例外：其斜率较为平缓，但得分不高，这是因为语料库的翻译质量不佳。因此，语料库的质量比数据集的规模更为重要。

6 结论与未来工作

本文介绍了使用统计机器翻译实现的15种印度语言与英语之间的双向机器翻译工作，并描述了这15种印度语言的语言特征。本研究还采用了一种量身定制的预处理方法。

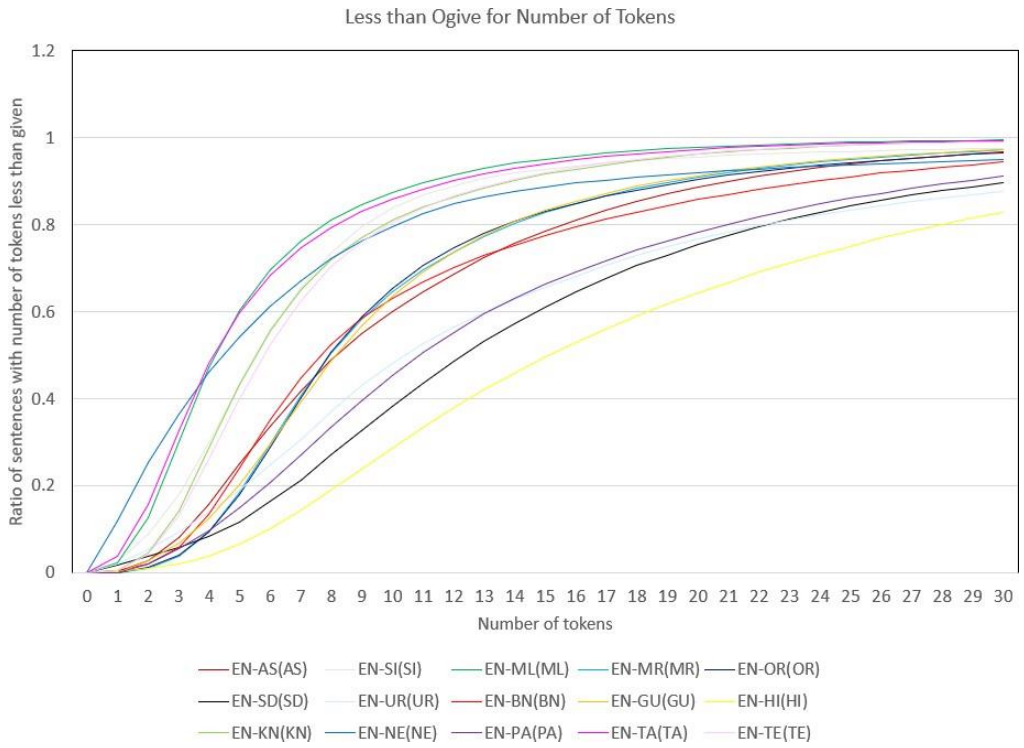

图1：全部十五个语言语料库中句子词元数的“小于”累积频数曲线

模型采用了 grow-diag-final-and 对齐模型和距离重排序模型。为评估翻译质量，本研究使用了 BLEU、RIBES 和 METEOR 等不同的评估指标。结果表明，所提出的统计机器翻译模型对于部分印度语言而言相当令人满意。然而，其性能水平与其他印度语言相比仍有差距，因此需要进一步改进。由于平行语料库数量有限且质量不高，所得指标值也相当低。

研究发现，部分语言的译文准确度不足。未来将探索用于验证语料库质量的方法，以评估语料库质量并剔除不准确的句行。达罗毗荼语通常是黏着语（词语由词素构成，每个词素都会为词语的含义作出贡献）。今后还将探索如何根据这些语言中词语的分解形式推导译文。

有趣的是，在某些印度语言中，微调方案正在影响质量。应通过降噪、语料清洗以及针对这些语言的微调方案等技术，分析并缓解这一现象的成因，以确保更高的质量。此外，还可以使用各种其他方法，对更多语言对和语料库进行分析与评估。统计与神经机器翻译混合系统，以及采用其他对齐模型和重排序模型等技术，也将作为后续研究的一部分加以探讨。

参考文献

Dorr, Bonnie J.、Eduard H. Hovy 和 Lori S. Levin。2004年。《自然语言处理与机器翻译：语言与语言学百科全书》。世界工程大会国际会议论文集，1–20。

Chapelle, Carol A. 和 Yoo-Ree Chung。2010年。《自然语言处理与机器翻译：语言与语言学百科全书》。《语言测试》，301–315。

So默斯，哈罗德。2011年。机器翻译：历史、发展与

局限。《牛津翻译研究手册》。

Tripathi, Sneha. 和 Juran Krishna Sarkhel。2010年。机器翻译方法。NISCAIR-CSIR，388-393

Lopez, Adam。2008年。统计机器翻译。ACM《计算调查》（CSUR），1-49

Stasimioti, Maria.，2020年。机器翻译质量：对统计机器翻译、神经机器翻译和定制神经机器翻译输出的比较评估。欧洲机器翻译协会第22届年度会议论文集，1-49

Michał Ziemski、Marcin Junczys-Dowmunt 和 Bruno Pouliquen.，2016年。联合国平行语料库 v1.0。载于第十届语言资源与评测国际会议论文集，3530–3534

Lohar, Pintu、Popović, Maja、Alfi, Haithem 和 Way, Andy.，2019年。统计机器翻译与神经机器翻译在翻译用户生成内容方面的系统性比较。计算语言学与智能文本处理国际会议，7-13

L, Zhou., Wenpeng Hu, J Zhang, C Zong., 2017年。用于机器翻译的神经系统组合。arXiv 预印本 arXiv:1704.06393，378–384

Wang, X., Zhengdong Lu, v, Hang Li, Deyi Xiong, Min Zhang., 2017年。由统计机器翻译指导的神经机器翻译。第三十一届 AAAI人工智能大会，1-7

Castilho, S., Moorkens, J., Gaspari,Federico., and Andy Way., 2017年. 神经机器翻译是否达到了最先进水平？《布拉格数学语言学通报》，109-120

Wang, Xing, Zhaopeng Tu, and Min Zhang., 2018. 将统计机器翻译的词汇知识融入神经机器翻译？IEEE/ACM《音频、语音与语言处理汇刊》，2255- 2266

Mahata, Sainik Kumar., Mandal,Soumil., Das, Dipankar., 和 Bandyopad-hyay, Sivaji., 2018。SMT 与 NMT：对印地语和孟加拉语简单句的比较。arXiv 预印本 arXiv:1812.04898

Koehn, Philipp.,Hoang, Hieu., Birch, Alexandra., Callison-Burch, Chris., Federico, Marcello., Bertoldi, Nicola., Cowan, Brooke., Shen, Wade., Moran, Christine., Zens, Richard., Dyer, Chris., Bojar, Ondřej., Con-stantin, Alexandra., 和 Herbst, Evan., 2018。Moses：统计机器翻译开源工具包。第45届计算语言学协会年会论文集——演示与海报会议论文集，2255-2266

Zens, Richard, Franz Josef Och, 和 Hermann Ney., 2002年。基于短语的统计机器翻译。人工智能年会，2255-2266

Hearne、Mary 和 Andy Way，2011年。《统计机器翻译：语言学家和译者指南》。《语言与语言学指南》，205-226

Costa-Jussà 和 Marta R.，2012年。《基于规则和统计的加泰罗尼亚语—西班牙语机器翻译系统的研究与比较》。《计算与信息学》，245-70

Carl, Michael 和 Andy, Way，2000年。迈向基于实例的机器翻译与基于规则的机器翻译的动态连接。《机器翻译》，223-257

Hutchins, W. John，2005年。迈向基于实例的机器翻译的定义。《基于实例的机器翻译研讨会》，223-257

Marton, Yuval., Chris Callison-Burch., 和 Philip Resnik., 2009年。利用单语派生的释义改进统计机器翻译。计算语言学协会，381–390

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, 和 P. S. Roossin., 1990年。一种机器翻译的统计方法。计算语言学，79–85

Stahlberg, Felix., 2020年。语言模型和层次模型在神经序列到序列预测中的作用。欧洲机器翻译协会，5–6

Ramesh, Akshai, Venkatesh Balavadhani Parthasarathy, Rejwanul Haque, 和 Andy Way., 2020 年。比较印地语到泰米尔语和英语到泰米尔语的统计机器翻译与神经机器翻译性能。Digital 1，86-102

Toral, Antonio, 和 Víctor M. Sánchez-Cartagena，2017年。对神经机器翻译与基于短语的机器翻译在9个语言方向上的多维评估。计算语言学协会，1063–1073

Dasgupta, Sajib、Abu Wasif 和 Sharmin Azam。“英语到孟加拉语机器翻译的最优方法”，2004年。英语到孟加拉语机器翻译的最优方法。国际计算机与信息会议，1-6

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Geor- gakopoulou, Y., Lohar, P., Way, A., Miceli Barone, A. V. 和 Gialama, M.，2017年。《使用专业译者对基于短语的统计机器翻译和神经机器翻译进行比较质量评估》。载于《机器翻译峰会论文集》

XVI，第116–131页

Islam Z、Tiedemann J、Eisele A.，2010年。英语到孟加拉语的基于短语机器翻译。欧洲机器翻译协会第14届年会论文集，1-8

Kaur、Jasleen 和 Gurpreet Singh Josan，2011年。从英语到旁遮普语转写的统计方法。《国际计算机科学与工程期刊》，1518-1527

Zbib、Rabih、Spyros Matsoukas、Richard Schwartz 和 John Makhoul，2010年。统计机器翻译中词汇平滑的决策树。计算语言学协会，428–437

Pushpananda、Randil、Ruvan Weerasinghe 和 Mahesan Niranjan，2014年。僧伽罗语-泰米尔语机器翻译：提升翻译质量。载于澳大拉西亚语言技术协会研讨会论文集，129-133

Nalluri, A. 和 Kommaluri, V.，2011年。使用 Joshua 的统计机器翻译：构建“enTel”系统的方法。印度语言句法分析，1-6

Ananthakrishnan Ramanathan、Jayprasad Hegde、Ritesh M. Shah、Pushpak Bhattacharyya 和 Sasikumar M.，2008年。简单的句法和形态处理有助于英语-印地语统计机器翻译。《第三届自然语言处理国际联合会议论文集》，1-6

Ramanathan、Ananthakrishnan、Hansraj Choudhary、Avishek Ghosh 和 Pushpak Bhattacharyya，2009年。格标记与形态学：解决英语-印地语统计机器翻译中流畅性问题的核心。在 ACL第47届年会联合会议暨

第4届 AFNLP国际自然语言处理联合会议论文集，800-808

Papineni, Kishore、Salim Roukos、Todd Ward 和 Wei-Jing Zhu，2002年。BLEU：一种机器翻译自动评价方法。载于计算语言学协会第40届年会论文集，311-318

Wołk, Krzysztof 和 Danijel Koržinek，2016年。复述质量评估自动评价指标的比较与适配。arXiv预印本 arXiv:1601.02789

Banerjee, Satanjeev 和 Alon Lavie，2005年。METEOR：一种改进了与人工判断相关性的机器翻译评估自动指标。载于机器翻译和/或摘要生成内在与外在评价指标ACL研讨会论文集，65-72

KM, Shiva Kumar、B. N. Namitha 和 R. Nithya，2015年。通用语料库与圣经文本语料库下英语到卡纳达语基线机器翻译系统的比较研究。《国际应用工程研究杂志》，30195-30201

Acharya, Praveen, 和 Bal Krishna Bal，2018。《统计机器翻译与神经机器翻译的比较研究：英语-尼泊尔语语言对案例研究》。《国际应用工程研究期刊》，90-93

Lita, Lucian Vlad、Abe Ittycheriah、Salim Roukos 和 Nanda Kambhatla，2003。《英语到孟加拉语神经机器翻译系统的开发》。载于《计算语言学协会第41届年会论文集》，152-159

Abujar, S., Masum, A. K. M., Bhattacharya, A., Dutta, S., and Hossain, S. A.，2021。使用全局

注意力机制进行英语到孟加拉语的神经机器翻译。《数据挖掘与信息安全新兴技术》，359–369

Li、Zhifei、Chris Callison-Burch、Chris Dyer、Sanjeev Khudanpur、Lane Schwartz、Wren Thornton、Jonathan Weese 和 Omar Zaidan，2009年。Joshua：用于基于句法分析的机器翻译的开源工具包。载于《第四届统计机器翻译研讨会论文集》，135-139

Ramesh、Gowtham、Sumanth Doddapaneni、Aravinth Bheemaraj、Mayank Jobanputra、Raghavan AK、Ajitesh Sharma、Sujit Sahoo 等，2022年。Samanantar：适用于 11 种印度语言的最大公开平行语料库集合。《计算语言学协会汇刊》，145-162

Tiedemann、Jörg，2012年。OPUS 中的平行数据、工具和接口。载于 Lrec，2214-2218

Costa-jussà、Marta R.、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi 等，2022年。不让任何语言掉队：扩展以人为中心的机器翻译。arXiv 预印本 arXiv:2207.04672，2214-2218

使用人数最多的200种语言有哪些，2022年。《Ethnologue：世界语言》https://www.ethnologue.com/guides/ethnologue200

Costa-jussà、Marta R.、James Cross、Onur Çelebi、Maha Elbayad、Kenneth Heafield、Kevin Heffernan、Elahe Kalbassi 等，2003年。《统计机器翻译中的最小错误率训练》。《第41届计算语言学协会年会论文集》，160–167

Khan, N. J., Anwar, W. 和 Durrani, N.，2017年。印度语言机器翻译方法综述。arXiv 预印本 arXiv:1701.04290

Lita, L. V., Ittycheriah, A., Roukos, S. 和 Kambhatla, N.，2003。大小写转换。载于第41届计算语言学协会年会论文集，152-159

Kumawat、Sudhakar 和 Nitish Chandra，2014年。英语到印地语统计机器翻译中的基于距离的重排序。《国际计算机应用杂志》，27-40

Koehn、P.、Hoang、H.、Birch、A.、Callison-Burch、C.、Federico、M.、Bertoldi、N.、Cowan、B.、Shen、W.、Moran、C.、Zens、R.、Dyer、C.、Bojar、O.、Constantin、A. 和 Herbst、E.，2007。Moses：统计机器翻译开源工具包。载于《计算语言学协会年会论文集》，177–180