使用递归神经网络和迁移学习的低资源神经机器翻译：英语到伊博语的案例研究

OCHEME ANTHONY EKLE*†‡, Tennessee Technological University, USA
BISWARUP DAS*†, Moscow Institute of Physics and Technology, Russia

在本研究中，我们开发了用于英语到伊博语翻译的神经机器翻译（NMT）和基于 Transformers 的迁移学习模型——一种在尼日利亚和西非超过 4000 万人使用的低资源非洲语言。我们的模型使用经过策划和基准化的数据集进行训练，该数据集从圣经语料库、当地新闻、维基百科文章和 Common Crawl 中编译而来，并由母语专家进行验证。我们利用递归神经网络（RNN）架构，包括长短期记忆（LSTM）和门控循环单元（GRU），并通过注意力机制进行增强，以提高翻译准确性。为了进一步提升性能，我们在 SimpleTransformers 框架中应用了 MarianNMT 预训练模型进行迁移学习。我们的基于 RNN 的系统取得了具有竞争力的结果，与现有的英语-伊博语基准接近。通过迁移学习，我们观察到性能提升了 +4.83 个 BLEU 点，达到大约 70 % 的翻译准确性。这些发现突显了结合 RNN 和迁移学习来解决低资源语言翻译任务中的性能差距的有效性。

CCS Concepts: • **Computing methodologies** → **Machine translation**; **Natural language processing**.

Additional Key Words and Phrases: Neural Machine Translation, Low-Resource Languages,English-to-Igbo Translation, RNN, LSTM, Natural Language Processing, Transfer learning

1 介绍

对于先进翻译系统的需求日益增长，以及全球范围内语言多样性的增加，在文本翻译方面带来了显著挑战。这一需求引发了自然语言处理（NLP）领域内的深入研究，特别是在机器翻译（MT）方面 [2] 。机器翻译是 NLP 的一个关键应用，涉及自动化系统在不同语言之间翻译文本。机器翻译的主要目标是确保目标语言中的翻译句子传达与源语言中原始句子相同的含义 [65] 。

过去的研究人员采用了多种方法来构建机器翻译系统，包括基于规则的机器翻译（RBMT）、基于语料库或统计的机器翻译（SMT），以及结合 RBMT 和 SMT 元素的混合机器翻译（HMT）。这些范式涵盖了基于单词、短语、树形结构的方法，以及最近的神经机器翻译（NMT）方法 [53] 。

神经机器翻译在机器翻译领域中成为一个重要的突破，利用深度神经网络和大型语言模型 [57] 更有效地进行翻译任务的建模 [65] 。通常，NMT 系统采用编码器-解码器架构 [65] ，其中编码器处理输入句子，解码器生成翻译输出 [18] 。这些架构，尤其是在增强了注意力机制时，展现了最先进的性能，并显著超过了早期的翻译技术 [3] 。

研究表明，与传统方法相比，神经机器翻译（NMT）模型不仅学习速度更快，而且产生的翻译更流畅、更准确。基于递归神经网络（RNN）的模型，尤其是那些使用长短期记忆（LSTM）和门控循环单元（GRU）结构的模型，由于其能够捕捉文本中的序列依赖性而被广泛使用。然而，尽管在 NMT 方面取得了进展，但针对低资源和濒危语言（尤其是非洲语言）的系统开发研究仍然有限 [26] 。为了解决这一空白，我们的研究集中于构建一个用于将英语翻译成伊博語的 NMT 模型，伊博語是一种在尼日利亚和西非地区有超过 4000 万人使用的主要非洲语言。

我们的主要贡献总结如下：

- 我们提出了首个高性能的低资源英语-伊博语翻译系统，结合了注意力机制 [3] 和全局注意力 [46] 。

*Both authors contributed equally to this research.
†This research was conducted while the author was at Moscow Institute of Physics and Technology (MIPT) , Russia.
‡Author is currently affiliated with Tennessee Technological University , USA.

Authors' addresses: Ocheme Anthony Ekle, oaekle42@tntech.edu, ekleanthony@phystech.edu, Tennessee Technological University, 1020 Stadium Drive, 406., Cookeville, TN, USA, 38505; Biswarup Das, Moscow Institute of Physics and Technology ， 9 Institutskiy pereulok,, Dolgoprudny, Moscow region, Russia, biswarup.das1@huawei.com.

- 我们整合了教师强制算法，以在训练期间为解码器提供准确的参考输出，从而提高翻译质量。
- 我们实施并比较贪婪解码和束搜索策略以提高翻译的流畅性和准确性。
- 我们使用大约 12,000 个平行句对进行了广泛的实验。我们的基于 transformer 的迁移学习模型 [50, 52, 78] 达到了 70 % 的最高翻译准确率，并在英语-伊博语翻译的现有基准上显著提高了 +4.83 BLEU 分。

本文的其余部分组织如下：2 节回顾了 NLP 和 NMT 相关文献。3 节概述了 RNN 的基本概念并讨论了其局限性。4 节介绍了我们提出的翻译模型的架构。5 节详细说明了实验设置，包括 RNN 变体、注意机制、解码策略、数据集的比较以及迁移学习的应用。6 节呈现了量化和质化结果以及性能评估。最后，6 和 7 节提供了全面的讨论、结论、识别出的局限性和未来研究方向。

2 相关工作

本节回顾了关于自然语言处理（NLP）、语言理解的层次、词语表示和机器翻译的文献。同时强调了基于图的学习和生成模型的最新发展，展示了它们如何与低资源语言翻译相关联。

2.1 自然语言处理和语言级别

自然语言处理 (NLP) 的研究始于 20 世纪 40 年代后期，结合了人工智能 (AI) 和语言学的概念，其中机器翻译 (MT) 是其最早的应用之一 [44, 49] 。1947 年，沃伦 · 韦弗提议使用密码学和信息理论进行语言翻译 [63]；然而，他的方法在词汇歧义方面遇到了困难。这个挑战激发了该领域更深入的探索。诺姆 · 乔姆斯基在 1957 年发表的标志性作品《句法结构》引入了生成语法，极大地影响了 NLP 和 MT 的发展 [43] 。这个时代也出现了语音识别、启发式推理和问答系统的早期努力。从 20 世纪 70 年代到 90 年代，NLP 随着语义解析和话语建模等创新而进步 [40] 。尚克和泰斯勒开发了语言理解程序，重点关注人类概念知识，如记忆和目标 [61] 。

自 20 世纪 90 年代以来，概率和数据驱动的方法主导了自然语言处理领域，出现了像概率解析和词嵌入这样的模型。计算硬件的进步进一步增强了语音和语言处理能力 [44] 。今天，研究人员专注于能够更好地处理语言变异性和歧义性的下一代应用程序。一个核心目标是设计出能够以人类方式学习和表示文本数据的计算技术。将符号化的文本转换为数字形式仍然是自然语言处理中的基础步骤。

语言层次：自然语言处理涉及多个语言层次：语音学、形态学、词汇、句法、语义、语用学和篇章 [40] 。这些层次是相互依赖且对于理解和生成语言至关重要的。歧义仍然是一个主要的挑战，通过使用基于知识的方法（例如，Mahesh 和 Nirenburg，1996）以及现代技术如 FastText [67] 来解决。语音学处理声音模式；形态学分析词的结构；句法检查句子的构造；词汇解释词的意义；语义学研究短语内部的含义；篇章处理文本单元间的连贯性；语用学评估意图和上下文意义 [44] 。

2.2 词向量表示

词向量表示在自然语言处理中对于捕捉词与词之间的语义关系至关重要 [12] 。早期的方法使用“一次性”编码，无法捕捉词之间的相似性。Firth 提出通过词的上下文来表示词，这为嵌入奠定了基础。Bengio 等人 [6] 引入了神经词嵌入，随后 Collobert 和 Weston [16] 应用了多任务和半监督学习。

Mikolov 等人 [48] 介绍了 Word2Vec，具有连续词袋（CBOW）和跳字模型架构。CBOW 在给定上下文的情况下预测一个词，而跳字模型则从给定的词预测上下文。GloVe 是另一个由 Pennington 等人 [55] 提出的广泛使用的预训练嵌入。Facebook 的 FastText [9] 捕获了子词信息和一词多义，后来的工作纳入了不确定性建模 [1] 。

最近，像 BERT [21] 、ELMo 和 GPT-2 [25] 这样的上下文化词嵌入已经超越了旧模型，胜过了单热编码和词袋方法 [30] 。这些嵌入现在在监督和非监督任务中都处于核心位置，为神经机器翻译、问答和信息检索提供动力。

在本研究中，我们采用 Keras tokenizer 通过基于形态学的分词技术将文本转换为数值序列。我们还使用通过 Keras 的嵌入层预训练的词嵌入来生成密集的向量表示。

2.3 神经机器翻译

机器翻译的起源可以追溯到沃伦·韦弗 1946 年提出使用密码技术进行语言翻译的建议 [36] 。在随后的几年里，研究人员开始结合基于规则和统计的方法，从而催生了用于神经机器翻译（NMT）的混合模型 [73] 。

现代 NMT 系统通常遵循序列到序列（Seq2Seq）范式，由编码器和解码器组成 [65] 。编码器将输入句子转换为一个固定长度的向量（隐藏状态），而解码器则基于该向量生成输出序列。然而，传统的编码器-解码器模型面临两个主要挑战:

(1) 单词之间的长距离依赖可能会降低翻译精度。Gusev 和 Oboturov 通过逆转编码器的方向来解决这个问题，以协助训练。(2) 将整个输入序列压缩到一个单独的向量中会引入信息瓶颈，特别是对于较长的输入。

Bahdanau 等人 [3] 使用注意力机制解决了这些挑战。注意力机制使解码器可以在每个步骤动态参考所有的编码器状态，而不是依赖于一个固定的向量，从而在解码过程中专注于相关的输入标记。这极大地提高了翻译质量，但该方法增加了计算复杂性，并且仍然在低资源数据和长期依赖性方面存在困难。

Luong 等人后来引入了局部注意力，以改进源标记和目标标记之间的对齐。Yang 等人和 Shaw 等人开发的自注意力网络进一步推进了序列建模，使得标记能够关注句子中的不同位置。

除了基于 RNN 的模型，如 LSTM 和 GRU 外，CNNs 也被用于研究 NMT。Kalchbrenner 等人 [37] 和 Bradbury 等人 [10] 应用了 CNNs 进行高效的序列建模，而 Gehring 等人 [27] 使用了带注意力的 CNNs 来建模长依赖关系。其他创新包括 Graves 等人提出的连接时序分类（CTC）[31] ，以及用于稳健预测的模型集成 [51] 。

神经机器翻译的突破来自 Vaswani 等人提出的 Transformer 架构 [68] 。该模型用自注意力取代了循环网络，实现了并行化和性能的提升。关键组件包括多头注意力和逐位置的前馈网络。Transformers 为 GPT-2 [25] 、BERT [32] 和 T5 [58] 等模型奠定了基础。

尽管它们取得了成功，基于 Transformer 的模型在低资源语言上表现通常不佳 [78] 。为了解决这个问题，提出了迁移学习：首先训练一个高资源（父）模型，然后将其知识转移到低资源（子）模型 [50, 52, 78] 。通过利用共享的表示，这种方法提高了性能。

2.4 图神经网络与自然语言处理的新兴趋势

虽然传统的 NLP 模型依赖于顺序或卷积架构，但最近的研究表明，图神经网络（GNNs）[4] 可以有效地建模语言数据中的结构关系，如句法和语义依赖 [69] 。GNNs 已被应用于诸如词嵌入 [22] 、知识图谱补全以及低资源机器翻译等任务中，通过捕捉图结构中的单词和短语之间的长距离交互。例如，在翻译任务中，将单词表示为节点，将它们的共现或句法关系表示为边，这使模型能够保留更丰富的上下文。

在平行研究中，基于 GNN 的技术亦已被证明对动态图流中的异常检测有效 [7, 13, 23, 24] 。这些贡献突出了 GNN 在结构化 NLP 和基于动态图学习中的日益重要的作用。

此外，大型语言模型（LLM）的出现，如 GPT、多语种 BERT 变体以及近期的生成嵌入技术，推动了翻译和多语言理解的边界。这些模型利用大规模的预训练来捕捉富含上下文的表示，并在低资源环境中实现强大的性能 [11, 72] 。生成嵌入越来越多地应用于跨语言检索、总结和神经翻译 [60, 70] ，提供了在整合上下文、基于图的和生成学习方面的新机会。

在本研究中，我们专注于英语到伊博语的翻译——一种资源匮乏的语言对。我们采用基于 RNN 的编码-解码框架，并通过注意力和全局注意力机制来增强它 [3, 46] 。我们还应用教师强迫算法 [41] 来指导训练过程中解码，并结合迁移学习以优化性能。

目标：通过利用这些技术，我们旨在构建一个有效的神经机器翻译模型，以弥合英语和伊博语之间的语言差距，为低资源自然语言处理应用的发展做出贡献。

3 背景和 RNN 架构

本节提供了循环神经网络（RNN）的基础背景，包括经典的 RNN 架构及其挑战。我们还探讨了诸如长短期记忆（LSTM）和门控循环单元（GRU）等高级变体。此外，我们介绍了序列到序列（Seq2Seq）编码解码模型，强调了它的局限性，并描述了注意机制如何解决这些挑战。

3.1 传统的 RNN 和梯度消失问题

循环神经网络（RNNs）通过维护隐藏状态来捕捉来自以前时间步骤的信息，从而对时间序列或顺序数据进行建模。这种能力使它们适合多种时间依赖任务，包括金融预测 [29] 、电力负荷预测 [17] 和水质监测 [47] 。

早期的 RNN 架构，如 Vanilla RNN、Elman 网络和 Jordan 网络 [56] ，由于梯度消失和梯度爆炸问题而遇到了限制。这些问题在通过时间的反向传播（BPTT）过程中出现，使得网络难以学习长期依赖 [19, 30] 。结果是，当时间步数显著增加时，传统 RNN 往往会有所挣扎。

图 1 展示了一个基本 RNN 的结构，其中同一组权重（U 、V 和 W ）在不同时间步中用于处理序列数据。

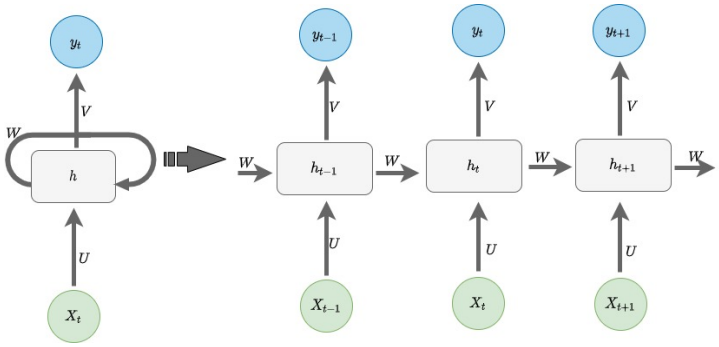


Fig. 1. **普通 RNN 架构。权重矩阵 U ， V 和 W 在时间步上共享以处理时间序列。**

为了克服这些挑战，引入了更先进的架构，例如长短时记忆网络（LSTM）和门控循环单元（GRU）[33] 。另一种值得注意的变体是连续时间 RNN（CTRNN），用于自适应行为建模和机器人技术 [5] 。这些改进使基于 RNN 的模型能够更有效地学习较长序列的时间依赖关系。

3.2 长短期记忆 (LSTM)

LSTM 网络由 Hochreiter 和 Schmidhuber 提出，旨在解决传统 RNN 中固有的梯度消失问题。LSTM 引入了门控机制，即输入门、遗忘门和输出门，以调节信息通过记忆单元的流动。这使得网络能够在较长时间间隔内保留或丢弃信息，从而非常适合于建模长期依赖。

LSTMs 在多个领域取得了最先进的成果，包括手写识别和语音处理 [62] 。它们还成功地使用连接时序分类（CTC）损失函数进行训练，用于需要灵活序列对齐的任务 [31] 。根据 Xuan Le [42] ，基于 LSTM 的模型在国际模式识别竞赛中设立了基准。

3.3 门控循环单元 (GRU)

门控循环单元（GRU）由 Cho 等人于 2014 年引入，是长短期记忆（LSTM）的一种精简替代方案。它将输入门和遗忘门的功能组合到一个更新门中，并使用重置门来控制先前隐藏状态的贡献。GRU 参数更少，提供计算效率，同时在许多自然语言处理应用中保持与 LSTM 相当的性能。

3.4 总结与实施

在本研究中，我们在编码器-解码器序列到序列框架中实现了 LSTM 和 GRU 架构。为了提高翻译精度，我们整合了注意力机制，该机制帮助模型在解码过程中关注输入的相关部分。此外，我们采用了教师强制算法，在训练期间使用真实输出来指导解码器。这些技术共同提高了模型处理长程依赖的能力，并提升了低资源语言对的翻译质量。

4 提出的方法

本节概述了用于构建和评估我们低资源的英语到伊博语神经机器翻译（NMT）系统的完整流程和方法。我们提供了关于数据获取、预处理、模型架构、训练和推断机制、迁移学习集成以及评估策略的全面解释。我们利用带有注意力机制的循环神经网络（RNNs），并探讨基于 Transformer 的预训练模型如何进一步提升翻译性能。

4.1 数据收集

大约 12,000 个英语-伊博语平行语料库句子是从 GitHub 存储库 中提取的，由 [26] 策划。这个基准数据集包括来自圣经语料库、当地新闻来源、维基百科文章、Common Crawl 数据和各种本土材料的经过验证的译文。语言专家对数据进行了审阅和清理，以确保对齐的一致性和语言的准确性。为了比较基准测试和迁移学习探索，我们还使用 Tatoeba 语料库中的英语-法语（Eng-Fra）数据集评估了我们的模型。

	English	Igbo
0	The news that will interest you:	Akụkọ ndị ga-amasị gị:
1	Joseph Achuzie, the Biafran brave man is gone.	Joseph Achuzie, Dike Biafra alala
2	Dapchi: Government has not defeated Boko Haram...	Dapchi: Gọọmenti emeribeghị Boko Haram - Massob
3	Tottenham look forward to lifting the FA cup i...	Tottenham na-ele anya iburu iko FA na mmeri Ro...
4	Son Heung-min, Fernando Llorente and Kyle Walt...	Son Heung-min, Fernando Llorente na Kyle Walte...
5	Son Heung-min of Tottenham and his team mates ...	Son Heung-min onye Tottenham na ndị otu ya na-...
6	Tottenham has won eight FA cup competitions.	Tottenham emerila n'asọmpị iko FA ugboro asatọ

Fig. 2. **样本英语-伊博语数据集可视化**

4.2 文本预处理和分析

为了确保原始的平行语料库可以用于训练，我们实施了一个三阶段的预处理流程：数据加载与可视化、分词和词汇构建。

4.2.1 **数据加载与采样**. 该平行语料库被加载到内存中，其中每个英语-伊博语句子对通过制表符隔开。这些句子对使用 pandas 数据框进行可视化，以验证一致性和格式，如图 2 所示。这个初始步骤帮助确保句子对齐是正确的，并适合于下游的分词和模型输入准备。

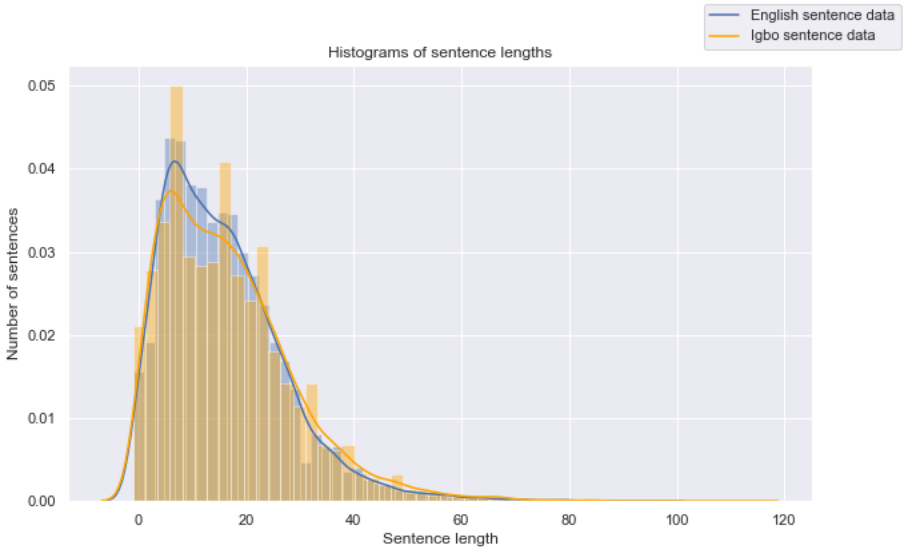


Fig. 3. **句子长度的直方图**

4.2.2 **句子长度分析**. 我们进行了探索性数据分析 (EDA)，以评估句子长度的分布。了解句子长度对于配置输入序列的填充和优化内存使用至关重要。如图 3 和 4 所示，大多数句子长度在 10 到 40 个标记之间。这些见解指导了填充策略，帮助我们避免过度的内存消耗或模型截断。

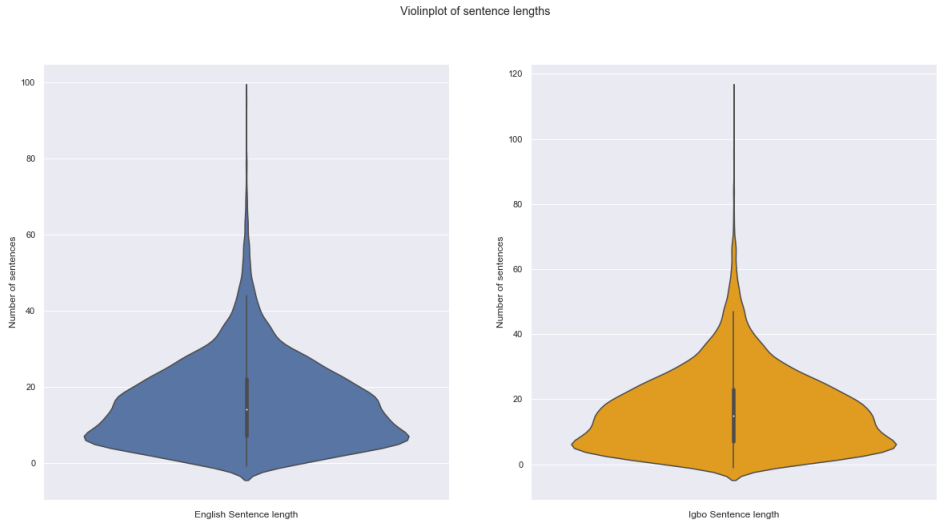


Fig. 4. **句子长度的小提琴图**

我们使用了 Keras 的 Tokenizer 进行形态识别的分词。标点符号和数字被过滤掉，所有文本都转换为小写。词元被映射为唯一的索引，以建立英语和伊博语的独立词汇表。特别的词元 <pad> 、<sos> 和 <eos> 分别被添加以表示填充、句子开始和句子结束。表格 1 展示了词汇表的大小和词数。

	English	Igbo
Total Words	176,375	184,538
Vocabulary Size	16,224	14,789

Table 1. **词汇统计**

4.3 模型设计：带注意力机制的 RNN

所提出的模型遵循经典的序列到序列（Seq2Seq）框架，并通过注意力机制进行增强。该架构由三个关键组件组成：一个编码器、一个解码器以及其间的注意力层。整体架构如图所示 5 。

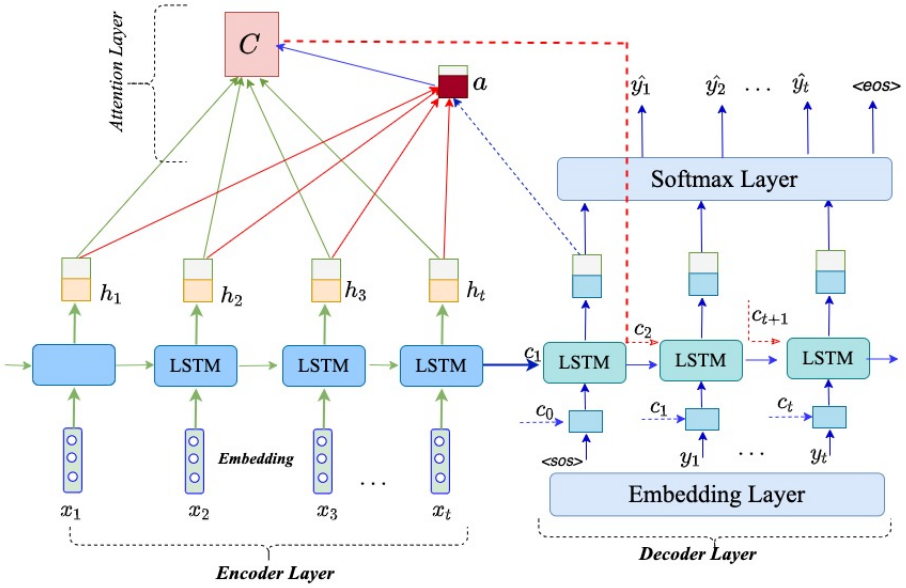


Fig. 5. **带有注意力机制的 RNN 编码器-解码器模型**

这种模块化结构使编码器能够处理和压缩输入序列为隐藏表示，注意力增强型解码器利用这些隐藏表示生成目标翻译。

4.3.1 **编码器** . 编码器使用一个嵌入层，后跟长短期记忆（LSTM）单元来处理输入序列。在每个时间步 t ，编码器计算隐状态 h_t 并更新上下文向量 C ：

$$h_t = \sigma_1(x_t, h_{t-1}), \quad C = \sigma_2(h_1, \ldots, h_t) \tag{1}$$

其中 σ_1 和 σ_2 是非线性激活函数。我们将编码器配置为批大小为 128，嵌入维度为 256，1024 个 LSTM 单元。根据验证性能和训练稳定性选择了这些超参数。

4.3.2 **带注意力的解码器**. 解码器镜像编码器的结构，并整合了 Bahdanau 风格的注意力 [3] 。在每个解码步骤中，解码器使用之前的输出和注意力来生成下一个词。注意力机制计算上下文向量 C_t 为：

$$C_t = \sum_j a_{tj} h_j \tag{2}$$

其中

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})}, \quad e_{tj} = \text{score}(h_{t-1}^{\text{dec}}, h_j^{\text{enc}}) \tag{3}$$

我们尝试了点积、一般、连接和缩放点积注意力。我们采用了全局注意力 [46] ，以其在可解释性和性能之间的平衡为由。

4.3.3 **激活函数和损失**. 编码器和解码器使用了 Sigmoid、Tanh 和 Softmax 激活函数。最终输出层应用 Softmax 来产生概率分布。我们使用稀疏 Softmax 交叉熵损失：

$$\mathcal{L}(y, f) = -\sum_{i=1}^{C} y_i \log\left(\frac{e^{f_y}}{\sum_{j=1}^{C} e^{f_j}}\right) \tag{4}$$

这个损失函数适用于多类分类，并在训练期间提供稳定的梯度。

Adam 优化器 [39] 使用了学习率 0.001。Adam 动态调整每个参数的学习率，结合了 AdaGrad 和 RMSProp 的优点，以确保在噪声环境中的收敛性和稳健性。

4.4 训练和推理

我们使用 TensorFlow 的 `GradientTape()` 构建了一个自定义训练循环，以记录操作和计算梯度。在训练期间，使用真实值标记对解码器应用了教师强迫。这种技术通过减少暴露偏差来加速收敛和提高翻译精度。

在推理时，解码器一次生成一个词，使用先前预测的词作为输入而不是真实值。图 6 可视化了训练和推理之间的区别。

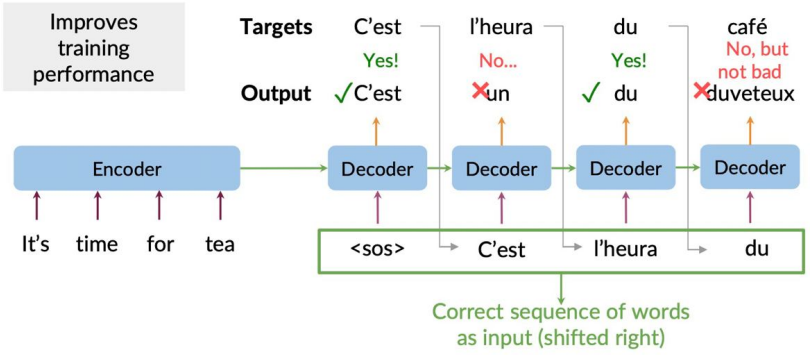


Fig. 6. **英文到法语翻译中的教师强制** [20]

解码策略：. 贪婪解码在每个步骤中选择最可能的标记：

$$\hat{y}_t = \arg\max_y P(y \mid x) \tag{5}$$

波束搜索解码保持多个假设（波束宽度为 2），如图 7 所示，通过探索更多的序列路径来提高翻译质量。

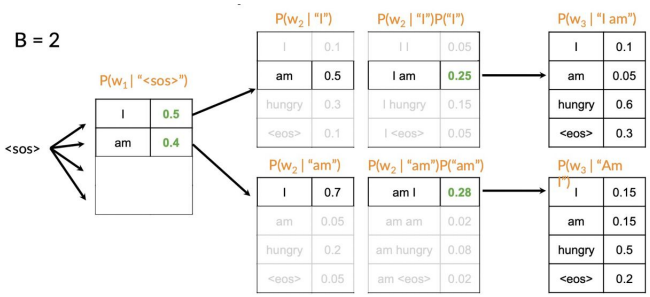


Fig. 7. **束搜索解码，其中束大小** = 2 [20]

为了评估翻译质量，我们使用了 BLEU 指标 [54]，它计算模型输出与参考翻译之间的 n-gram 重叠。我们测试了 300 个英语-伊博语句子对。BLEU 分数范围从 0（差）到 1（完美匹配）。我们的 RNN 模型相较于之前的基准提升了 +4.83 的 BLEU 分数，达到了 70% 的翻译质量。

4.5 迁移学习

为了进一步提高性能，我们使用基于 MarianNMT 的 Simple Transformers 库进行了迁移学习。迁移学习在资源匮乏的环境中特别有价值。HuggingFace 提供了几十种预训练模型，而 MarianNMT 由 Microsoft Translator 团队开发，经过优化可实现快速神经翻译。它由在 8 个 GPU 上使用同步 Adam 训练的 6 层编码器-解码器堆栈组成。这种模型支持多种语言对，包括英语-伊博语。

通过在我们的平行数据上微调 MarianNMT，我们从之前学习过的语言表示中获益，提升了泛化能力和翻译流畅性。预训练的权重作为一种基础，减少了我们模型从零开始学习的负担。

总之，我们提出的方法表明，将经典的 RNN 与现代注意力技术和迁移学习相结合可在资源稀缺的翻译任务中获得强大的性能。该流程是模块化和可扩展的，允许未来与 transformer 编码器、多语言嵌入和元学习框架进行集成。

5 实验与结果

本节展示了我们的神经机器翻译模型的实验结果。我们评估了模型结构（LSTM 对比 GRU）、注意力评分函数、解码策略和超参数调整。实验在英语-伊博语平行语料库上进行，并在英语-法语数据集上进行了额外的验证。性能通过 BLEU 分数、损失值、执行时间和定性翻译示例进行衡量。

5.1 实验设置

数据集：我们使用了一份精选的英语-伊博语平行语料库，其中包括从多个来源提取的大约 12,000 对句子对，这些来源包括《圣经》文本、维基百科、当地新闻和 Common Crawl。数据经过母语专家的预处理、分词和验证。为了进行附加基准测试，我们使用了来自 Tatoeba 多语言集合的 70,000 对英语-法语句子对。

系统配置：所有训练和评估实验均在 Google Colab 上使用 Tesla K80 GPU 进行。初始模型开发、可视化和预处理

是在一台个人的 MacBook Pro（macOS 13.2，16GB RAM，Apple M1）上进行的。深度学习模型是在 TensorFlow 2.x 中使用 Keras 实现的。

5.2 模型性能：LSTM 与 GRU 的比较

我们使用相同的超参数设置实现并评估了基于 LSTM 和 GRU 的 Seq2Seq 模型。表 2 提供了模型配置、执行指标和性能结果的全面摘要，而图 8 对比了它们在 80 个周期内的训练损失和 BLEU 得分轨迹。

尽管 GRU 模型的训练速度大约比 LSTM 模型快 2 倍（2.53 小时对比 5.25 小时），并且 BLEU 分数高出 2.8 个百分点（0.36 对比 0.35），但 LSTM 在训练损失和对较长及更复杂句子的泛化上始终表现得更好。这些观察结果在图 8 中得到了直观的强化。

对样本翻译的定性分析（表 3 和 4 ）显示，与 GRU 相比，LSTM 模型生成了更流畅且在上下文中更准确的输出。鉴于我们的数据集的性质——其中包含大量的长句——我们选择 LSTM 架构作为最终的部署模型。

Table 2. LSTM 和 GRU 架构的比较。**在使用相同超参数对英语-伊博语翻译进行评估时，GRU 在短序列上训练更快且表现更好，而 LSTM 在较长输入上产生较低的损失并提供更流畅的翻译。**

Metric	LSTM	GRU
Units	1024	1024
Embedding Dimension	256	256
Batch Size	128	128
Epochs	80	80
Time per Epoch	~ 3.88 mins	~ 1.88 mins
Total Exec. Time	~ 5.25 hrs	~ 2.53 hrs
Mean Loss	0.0144	0.0258
BLEU Score	0.350	0.360

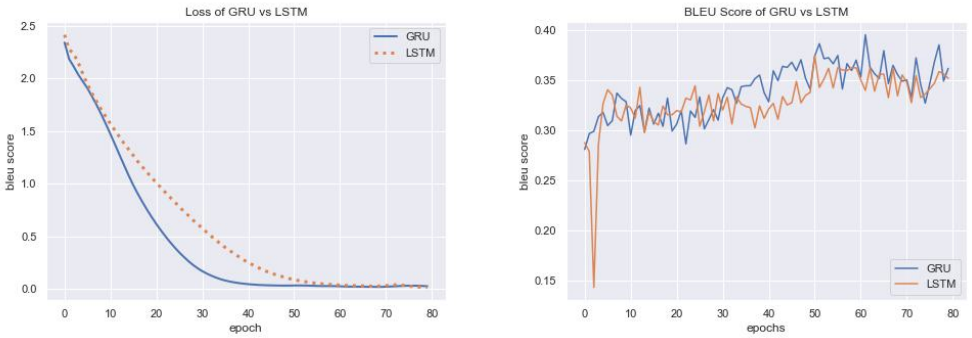


Fig. 8. GRU 与 LSTM 模型的损失和 BLEU 分数比较。**在 80 个周期中显示训练损失和 BLEU 分数的变化趋势。虽然 GRU 收敛速度更快，但 LSTM 在一段时间内表现出较低的损失和更平稳的翻译准确性。**

英语-伊博语翻译示例（LSTM 模型）

Source Sentence	interesting news
Actual Translation	akụkọ na-atọ ụtọ
Predicted	akụụkọ ndị ga amasị gị 含义： 你喜欢的故事
Source Sentence	the Nigerian government has also asked the South African government to pay for damages done to Nigerian businesses in the country.
Actual Translation	Gọọmentị nke Naịjirịa ... mmebị ha mere n ' azụmaahịa nke ala Naịjirịa
Predicted	gọọmentị naijirịa ... gọọmentị swiss weghachitere ga eme. 意义： 尼日利亚政府已告知尼日利亚政府，瑞士政府已经返回。
Source Sentence	Super Eagles is interesting.
Actual Translation	super ugọ na-adọrọ mmasị.
Predicted	ndị ọtu Super Eagles. 意思： 超级雄鹰队成员
Source Sentence	No money was brought for handover in Imo State.
Actual Translation	Enweghị ego enyere maka mbufe na agụmakwụkwọ steeti.
Predicted	ewepụtaghị ego maka nyefe ọchịchị Imo steeti 意义： 尚未拨出资金用于州政府的转移
Source Sentence	Commissioner for education Mr. Maazi Oladoyin Folorunsho, gave the order while inspecting the school to assess the level of damage by the students.
Actual Translation	Kọmishọna nke Agụmakwụkwọ ... mmebị ha mere.
Predicted	kọmishọna na ahụ maka ọrụ maazị marcel ifejiofor ... 意义： 专员马塞尔 · 伊斐吉奥弗称赞了地区建设委员会的努力。涉及该地区……

Table 3. **来自 LSTM 模型的示例翻译。对于每个例子，我们显示输入句子、实际翻译和预测翻译。值得注意的标记用粗体显示，并提供 预测输出的含义 以便于解释。**

English-Igbo Translations

src sentence	she was born in..
actual translation	a mụrụ ya n afọ.
predicted	a mụrụ ya n afọ. **meaning:** *she was born.*
src sentence	pillars Abia warriors have resolved the feud over enaholo
actual translation	pillars abịa warriors emeziela ihu mgbarụ dị n ' etiti ha maka enaholo.
predicted	pillars abịa warriors ekpeziela esemokwu banyere enaholo. **meaning:** *pillars Abia warriors have settled the dispute over enaholo.*
src sentence	super eagles is interesting
actual translation	super ugọ na-adọrọ mmasị.
predicted	ndị ọtu super eagles. **meaning:** *members of the super eagles*
src sentence	he now holds the position of minister of state agriculture.
actual translation	ugbua o ji ọkwa minista na ahụ maka ọrụ ugbo nke steeti.
predicted	o jizi ọkwa dịka ọtụtụ ndị ndọrọndọrọ ọchịchị... **meaning:** *he holds the position of as many politicians...*
src sentence	california fire
actual translation	ọkụ kalifornia
predicted	ọkụ kalifonia **meaning:** *California fire*

Table 4. **来自 GRU 模型的示例翻译：对于每个例子，我们显示预测的翻译以及实际的 Igbo 参考。显著的标记用粗体强调，预测句子的含义包含在 蓝色 中。**

5.3 注意力评分函数的评估

为了增强解码器的上下文感知能力，我们评估了三种广泛使用的注意力评分函数：Bahdanau 等人提出的基于连接的方法 [3]，以及 Luong 等人引入的点积和通用评分机制 [46]。这些函数在全球注意力框架中实现，并在相同的模型配置下进行了测试。

表 5 总结了不同注意力类型的结果。其中，点积注意力在翻译质量和计算效率之间达到了最佳平衡，得到了最高的 BLEU 分数，同时保持了每个 epoch 最低的执行时间。尽管连接方法导致训练损失略低，但由于额外的参数化，计算成本显著增加。

图 9 绘制了每种注意力变体在历经各个时期的训练损失和 BLEU 分数的进展，清晰地展示了点积方法的良好收敛行为。此外，我们在表 6 和表 7 中展示了定性结果，展示了输出样本及其在不同注意力机制下的语义对齐。

为了进一步分析模型行为，我们可视化了由每个评分函数生成的对齐权重，突出展示了点积注意力如何始终生成更清晰、更集中的对齐映射。基于定量和定性的评估，点积注意力被选择用于整合到最终的模型架构中。

与 LSTM 和 GRU 基线的比较。与第 5.2 节中的基线模型相比，单独的 GRU 在 BLEU 得分上稍稍优于 LSTM（0.36 对 0.35），但损失更高，整合注意力—特别是点积注意力—导致了更显著且一致的性能提升。LSTM+ 点积模型不仅保留了 LSTM 在长序列上的鲁棒性，还在训练时间几乎减少 50 %（从 ~ 5.25 小时到 ~ 3 小时）的同时实现了 +0.01 的 BLEU 提升。

表 6 和 7 中的定性样本进一步说明了点积注意力在短句和长句中均能产生更流畅且语义一致的翻译。这些改进证实了注意力机制的引入在低资源神经翻译任务中提供了有意义的增益，使其成为最终系统架构的宝贵补充。

Table 5. **不同注意力评分函数的性能** - LSTM + **点积模型的训练速度快** 50 %，**并且与连接和一般评分相比，获得了更高的** BLEU **分数**。

Metric	Concatenation	Dot-product	General
Execution Time per Epoch	~ 4 mins	~ 2 mins	~ 3.8 mins
Total Execution Time	~ 5.64 hrs	~ 3 hrs	~ 5 hrs
Mean Loss	0.0303	0.1010	0.8421
BLEU Score	0.358	0.36	0.315

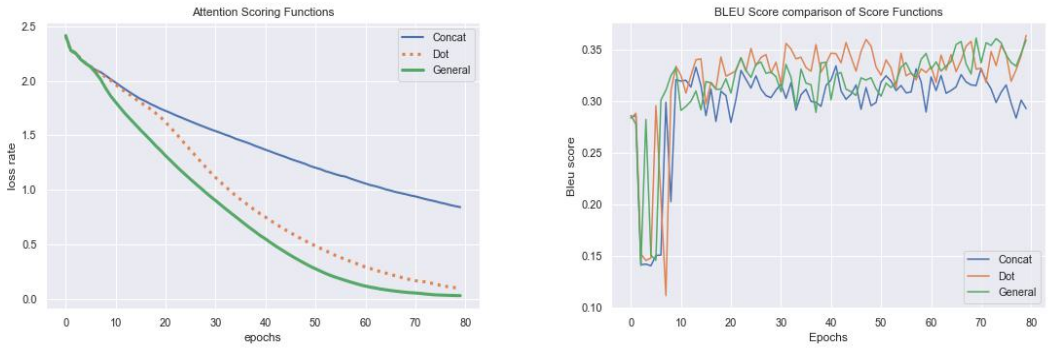


Fig. 9. **注意力评分函数的损失和** BLEU **分数。比较在** LSTM **模型中使用不同全局注意力评分机制（连接、点积和通用）的训练损失和** BLEU **分数**。

English-Igbo Translations	
src sentence	interesting news
actual translation	akụkọ na-atọ ụtọ
Predicted (**fairly close**)	akụkọ ndị ga amasị gị **meaning:** *story you like*
src sentence	no money was brought for handover in Imo State.
actual translation	Enweghị ego enyere maka mbufe na agụmakwụkwọ steeti.
Predicted (**fairly close**)	ewepụtaghị ego n imo steeti **meaning:** *no budget for state education*
src sentence	Commissioner for Education Mr. Maazi Oladoyin Folorunsho gave the order while inspecting the school to assess the level of damage by the students.
actual translation	Kọmishọna nke Agụmakwụkwọ, Maazị Oladoyin Folorunsho, nyere iwu a mgbe ọ na-eme nnyocha maka ihe niile nke ụmụakwụkwọ ahụ mebiri.
Predicted (**fair**)	gọvanọ nke a ga eme ka ọ bụ mazi uchenna okafor kwuru na ndị otu ahụ kwuru... **meaning:** *the governor will act, Mr. Uchenna Okafor said...*
src sentence	One of the greatest dangers for any business is the violation of the rule of law and the principles of justice, equality, and the right to be heard.
actual translation	Otu n'ime ihe kasị dị egwu maka azụmahịa ọbụla nke na-ada iwu ọchịchị na usoro ikpe mkwụmọtọ, ịha ụha na ikike ịnụ olu.
Predicted (**fair**)	iwu dị egwu n azụmaahịa maka ndị niile **meaning:** *a dangerous law in business for everyone*

Table 6. **通用评分函数的样本翻译：我们展示了使用通用注意力翻译的精选例子。每项内容包括源句子、参考翻译、模型预测以及在 蓝色 中的解读意义。标记为 fairly close 的结果显示了在意义和结构上的适度对齐。**

English-Igbo Translations	
src sentence	important news today
actual translation	akụkọ na-atọ ụtọ
Predicted (**good**)	akụụkọ dị mkpa taa **meaning:** *important news today*
src sentence	no money was brought for handover in Imo State.
actual translation	Enweghị ego enyere maka mbufe na agụmakwụkwọ steeti.
Predicted	ewepụtaghị ego maka nyefe ọchịchị imo steeti **meaning:** *has not set aside funds for the transfer of state government*
src sentence	participants at a two-day capacity training workshop in Abakaliki, Ebonyi State, have identified the deconstruction of patriarchal dominant views in society as a major step towards actualizing thirty-five percent affirmative action.
actual translation	ndị sonyere na nkụzị ụbọchị abụọ n abakiliki ebonyi steeti achọpụtala nruzighari nke echiche ndị isi ala na ọha mmadụ dịka nnukwu nzọụkwụ iji gosipụta mmezu nkwupụta pesent iri atọ na ise.
Predicted (**good on long sen.**)	ndị sonyere na nkụzị ụbọchị abụọ n abakiliki ebonyi steeti achọpụtala nruzighari nke echiche ndị isi ala na ọha mmadụ dị ka nnukwu nzọụkwụ iji gosipụta mmezu nkwupụta iri atọ na ise. **meaning:** *Participants in the two-day program in Ebonyi identified dismantling dominant views as key to 35 % affirmative action.*
src sentence	the warriors who will compete in Imo and Abia states
actual translation	dike ndị ga akwata ya n imo na abia steeti
Predicted (**very good**)	dike ndị ga akwata ya n imo na abia steeti **meaning:** *the warriors who will compete in Imo and Abia states*
src sentence	He said that the law, if enacted, would protect, respect, and promote the rights of women and children to improve their health and standard of living.
actual translation	o kwuru na ọ bụrụ na e mebe iwu ahụ na ọ ga echekwa sopụrụ ma kwalite ikike dịịrị ụmụnwaanyị na ụmụaka iji bulite ahụ ike ha nakwa obibi ndụ ha.
Predicted (**acceptable**)	o kwuru na iwu ahụ ga echekwa ikike ụmụnwaanyị na ụmụaka... **meaning:** *he said the law will protect the rights of women and children...*

Table 7. **点积注意力样本翻译：对于每个样本，我们展示原始输入句子、参考翻译和模型预测的输出。关键的预测短语在上下文中用其在 蓝色 中的解释意义进行高亮，以展示语义忠诚度。这包括对短句和长句的评估。**

5.4 解码策略与超参数优化

为了进一步提高翻译质量，我们尝试了不同的解码策略，并调整了关键的超参数以实现最佳的模型配置。

解码技术。我们比较了贪婪解码和集束搜索（束宽为 5），以评估它们对翻译准确性的影响。集束搜索在短句子的流畅性上略有改善。然而，对于较长且语法更复杂的句子，贪婪解码始终产生更好的结果。表格 8 显示了使用两种方法生成的示例输出，揭示了贪婪解码在具有挑战性的序列中更好地保持了语义对齐。因此，贪婪解码因其优越的泛化能力和简单性被采用于我们的最终系统。

English-Igbo Translations: Greedy Algorithm vs. Beam Search	
Source Sentence	interesting news
Actual Translation	akụkọ na-atọ ụtọ
Predicted (Greedy)	akụkọ ndị ga amasị gị **meaning:** *story you like*
Predicted (Beam)	akụkọ dị mkpa **meaning:** *interesting story*

Table 8. **贪心解码与集束搜索的比较：虽然这两种方法都能产生有效的翻译，但在这种情况下，贪心解码会产生更个性化的结果。**

超参数调优。我们进行了系统实验，以微调模型的批量大小和 dropout 率。如图 10 所示，批量大小为 32 和 dropout 为 0.5 提供了收敛稳定性和泛化之间的最佳折衷，最大限度地减少训练损失并在多次运行中提升 BLEU 分数。

我们还跟踪了训练过程中误差的减少。表格 10 显示，经过 50 次迭代，模型实现了平均损失 0.2727 和 BLEU 分数 0.3303 。经过 100 次迭代后，这一情况显著改善，最终损失达到 0.0143 ， BLEU 得分为 0.3817 ，反映出较强的学习曲线。

最终模型表现。表格 9 展示了所有配置的综合比较。我们的最终模型——使用 LSTM、点积注意力、贪心解码以及优化的超参数构建——获得了最高的 BLEU 分数 0.3817，超越了所有先前设置，并在公开的英语-伊博语基准测试中表现优于其他模型（表格 11 ）。

Table 9. **性能比较总结——不同模型配置下的 BLEU 分数和训练见解。最终模型整合了所有改进，显示出卓越的性能。**

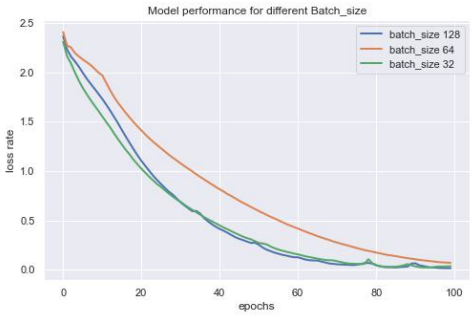
Model Variant	BLEU Score	Loss	Time (hrs)
GRU (Baseline)	0.360	0.0258	2.53
LSTM (Baseline)	0.350	0.0144	5.25
LSTM + General Attention	0.315	0.8421	5.00
LSTM + Concat Attention	0.358	0.0303	5.64
LSTM + Dot Attention	0.360	0.1010	3.00
Final Optimized Model	0.3817	0.0143	2.85

Table 10. **模型在训练阶段的表现：比较 100 轮次训练周期的前 50 次和后 50 次的平均错误和 BLEU 分数。模型在后半段显示出显著的改进。**

Epoch Range	Mean Error	BLEU Score
First 50 Epochs	0.2727	0.3303
Last 50 Epochs	0.0143	0.3817

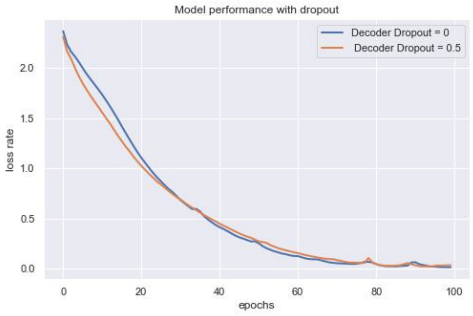
Table 11. **英语-伊博语数据集上的基准表现：JW300.en.ig 和 Tatoeba.en.ig 测试集的 BLEU 和 chr-F 分数，基于 Helsinki-NLP 开发并托管在 Hugging Face [34] 上的预训练翻译模型。**

HuggingFace Benchmarks	BLEU Score	Character F-score
JW300.en.ig	39.5	0.546
Tatoeba.en.ig	3.8	0.297

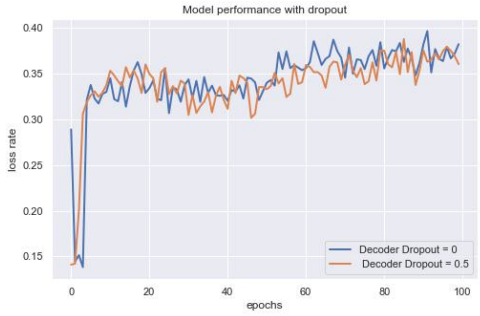


(a) Loss across Batch Sizes

(b) BLEU Score across Batch Sizes



(c) Loss across Dropout Rates



(d) BLEU Score across Dropout Rates

Fig. 10. **最终优化模型性能：在不同超参数配置下的损失和 BLEU 分数比较。最佳性能是在批量大小为 32 和 dropout 为 0.5 时实现的，这导致了更平稳的训练动态和提高的翻译准确性。**

为了进一步评估我们模型的泛化能力，我们在英法 Tatoeba 语料库的一个包含 70,000 个句子的子集上训练了它。表 12 中展示的结果表明，串联和一般注意力评分函数取得了最佳性能，分别获得了 0.590 和 0.587 的 BLEU 分数。这些分数超过了官方 Tatoeba 基准的 0.505 BLEU 分数。

这种改进可能部分归因于英语和法语之间的句法相似性，两者都遵循主-谓-宾（SVO）结构。我们特别关注这种简化的语言对齐，以测试我们的模型在更有利的结构条件下的表现。

该模型训练了 80 个周期，使用了不同的注意力打分机制。如表 12 所示，串联和一般打分方法分别达到了 0.590 和 0.587 的 BLEU 分数，相应的损失值为 0.0620 和 0.0818。相比之下，虽然点积注意力在计算上更高效——训练速度比一般注意力快约 82%，比串联快 56%——但其产生的 BLEU 分数较低，仅为 0.526。

这一结果表明点积注意力可能无法很好地在英法 Tatoeba 子集上推广。需要进一步分析以调查其相对较弱的语义对齐。

我们的研究结果表明，尽管收敛速度较慢，通用和串联注意力方法在该数据集的翻译质量上优于点积。这些 BLEU 分数也超过了赫尔辛基-NLP OPUS-MT 模型 [66] 报告的基线得分 0.505，表明我们方法的强大之处。

Table 12. **在英法数据集上的表现：使用 Tatoeba Eng-Fra 语料库的一个子集比较注意力评分函数。在 BLEU 得分和平均损失上，concat 和 general 都优于点积方法。**

Metric	Concat	Dot-Product	General
Time per Epoch	~ 1.76 mins	~ 1.33 mins	~ 3.17 mins
Total Exec. Time	~ 2.36 hrs	~ 1.80 hrs	~ 4.30 hrs
Mean Loss	0.0620	0.2480	0.0818
BLEU Score	0.590	0.526	0.587

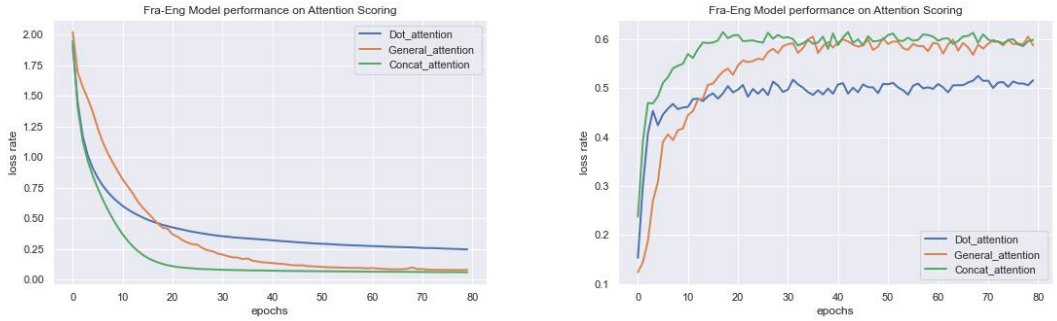


Fig. 11. **法语-英语数据集的损失和 BLEU 分数: 比较使用不同注意力评分机制的训练损失和 BLEU 分数随时期变化的图表。**

5.5 迁移学习的效果：在英语-伊博语上的微调

为了评估迁移学习的影响，我们使用 `SimpleTransformers` 框架对预训练的 MarianNMT 模型 [59, 71] 进行了微调，针对我们的英语-伊博语数据集。训练在 597 个句子对上进行了 20 个 epoch，使用的是 NVIDIA Tesla K80 GPU。每个 epoch 大约需要 15 分钟，总共用时 5.13 小时。

5.5.1 **训练和评估结果**. 该模型取得了 0.43 的 BLEU 分数，超越了 HuggingFace 的基准，分别为 0.38（Tatoeba.en.ig）和 0.395（JW300.en.ig）。图 12 显示了 BLEU 得分的分布：27 % 的翻译得分低于 0.35，40 % 介于 0.35 至 0.50 之间，33 % 超过了 0.50。

5.5.2 **训练动态和 *GPU* 时间**. 最终的训练损失为 0.6020，通过 `WandB` [8] 追踪。如图 13 所示，损失一直在下降，表明模型尚未完全收敛。通过更多的训练时间（额外的 10–20 个 epochs），预计会有进一步的改进。

Table 13. **BLEU 分数与词汇量比较：在 HuggingFace 基准模型和我们开发的模型中评估翻译性能。迁移学习方法（MarianNMT）产生最高的 BLEU 评分，而我们最终优化的 RNN 模型在词汇量更精简的情况下实现了可比的结果。**

Model	BLEU Score	Vocabulary Size
HuggingFace (Tatoeba)	0.380	18K
HuggingFace (JW300)	0.395	20K
Final Optimized RNN-based (Ours)	0.3817	16K
Transfer Learning (Ours)	0.4300	16K

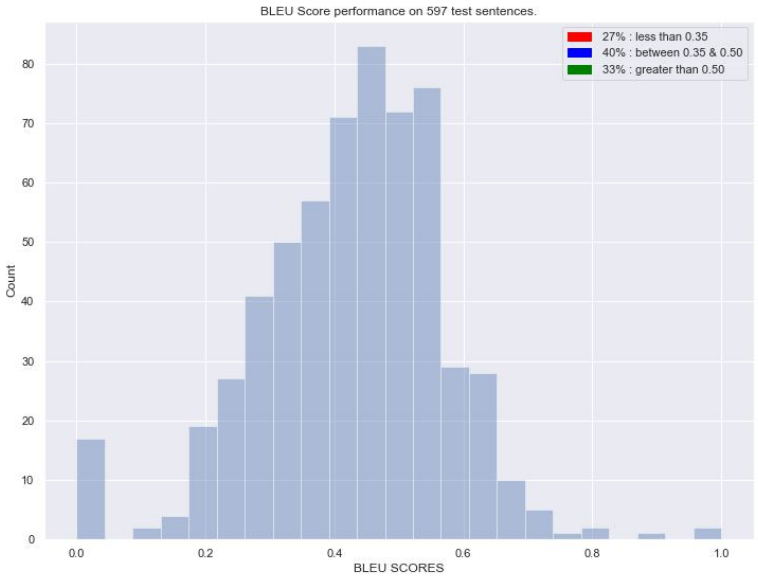
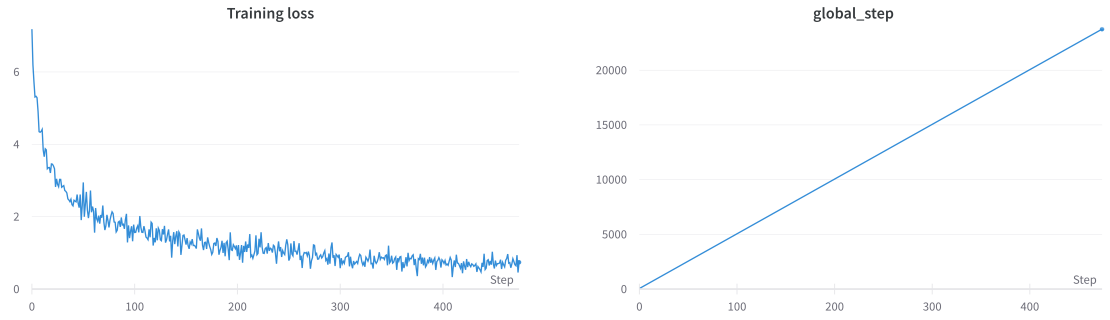


Fig. 12. BLEU **评分表现：使用微调后的** MarianNMT **模型在** 597 **个测试样本中的** BLEU **评分分布。**



(a) Loss trend over 20 epochs (b) Time taken per global step

Fig. 13. MarianNMT **微调的训练进展：这些图表展示了在英语到伊博语数据集进行微调时的关键动态变化。图** (a) **显示了在** 20 **个周期中训练损失的稳定下降，表明模型仍在学习且尚未完全收敛。图** (b) **展示了每个全局训练步骤所需的时间，表明在整个运行过程中** GPU **利用率的一致性和训练的稳定性。这些趋势支持通过更长时间的训练实现更好性能的潜力。**

5.5.3 **与基准和基于** *RNN* **的模型的比较**. 表格 13 将我们的微调模型与基于 RNN 的系统和现有的 HuggingFace 基准进行了比较。我们的迁移学习模型在保持紧凑词汇量的同时，达到了最高的 BLEU 得分 0.43。

最后，表格 **??** 给出了迁移学习模型的定性例子。它在常规、长、以及特殊字符的句子中表现出色，在超过 70 % 的情况下实现了准确且流畅的翻译。

6 讨论

本节综合了我们实验中的关键见解，涵盖了架构选择、注意力机制、解码策略、对其他语言对的泛化，以及迁移学习的效果。

6.1 模型架构：LSTM 与 GRU

我们的结果表明，基于 GRU 的模型训练速度更快，在短句上表现略好，取得了 0.36 的 BLEU 分数，而 LSTM 则为 0.35。然而，LSTM 模型在较长且句法复杂的句子上始终优于 GRU，表现为更低的损失（0.0144 对 0.0258）和更稳定的翻译准确性。这些发现与 Khandelwal et al. [38] 一致，促使我们在最终模型中采用了 LSTM 层。

我们在 Luong 的全局注意力框架下实现并比较了三种注意力评分策略——串联法、点积法和通用法。点积法在翻译质量和计算效率之间提供了最佳的折衷。尽管串联法的损失略低，但点积注意力训练速度快近 50%，并产生了更流畅的翻译，证实了 Bahdanau et al. [3] 和 Luong et al. [46] 的结论。

我们评估了贪婪解码和束搜索（束大小 = 5）。令人惊讶的是，贪婪解码整体表现更好，尤其是在长句子和上下文复杂的句子上，这一点也被 Sutskever et al. [65] 注意到了。此外，超参数调整显示，批量大小为 32，dropout 为 0.5，使用 1024 个 LSTM 单元和 100 个训练回合能够产生最佳性能，达到了 0.3817 的 BLEU 得分和 0.0143 的最终损失。这些结果优于 Tatoeba 基线（0.38）并接近 JW300 基准（0.395）。

为了评估模型的泛化能力，我们将其应用于一个来自 Tatoeba 的结构较为简单的英语-法语数据集，该数据集具有 SVO 结构。在 70,000 个句子对上，串联和一般评分函数的 BLEU 得分分别达到 0.59 和 0.587，超过了官方基准 0.505。这表明我们的架构在具备不同形态和句法复杂度的语言对之间具有适应性。

通过 SimpleTransformers 框架微调预训练的 MarianNMT 模型，在我们的英语-伊博语测试集上取得了 0.43 的 BLEU 分数。这一性能显著超过了我们基于 RNN 的模型（0.3817）和 HuggingFace 基准（Tatoeba: 0.38，JW300: 0.395），展示了迁移学习在低资源机器翻译中的强大优势。如表格 13 和图 12 所示，这一方法改善了在不同句子结构中的泛化能力、流畅性和语义保留。

总体而言，我们的实验强调了一种经过良好调优的基于 LSTM 的架构，结合点积注意力和优化的超参数，即使在资源匮乏的翻译任务中也能表现出色。迁移学习进一步提升了性能，使我们的模型成为英语-伊博语和类似缺乏代表语言对的强力基线。这些发现与更广泛的研究一致，例如 Chen et al. [14]，强调了传统的基于 RNN 的架构在适当优化后依然是序列到序列任务中极为有效的选择。

7 结论

在这项研究中，我们开发了一种基于循环神经网络（RNN）的神经机器翻译（NMT）系统，用于英语到伊博语的翻译，解决低资源语言任务。我们的实验强调了结合注意力机制，尤其是点积注意力的长短期记忆（LSTM）模型在产生流畅且语义准确的翻译方面的有效性。尽管在有限的数据上进行了训练，我们的模型实现了与最新基准如 Tatoeba 和 JW300 相当的性能。

此外，通过将我们的实验扩展到英语-法语数据集，我们展示了我们架构的普遍适用性。该模型在 Tatoeba 英语-法语基准测试中超出了 9 个 BLEU 点。这种跨语言的表现强调了我们方法在不同语法复杂性的语言中表现出的稳健性。

为了进一步提高翻译质量，我们使用 MarianNMT 框架进行了迁移学习。这产生了 0.43 的 BLEU 分数，比现有的 HuggingFace 英语-伊博语基线（Tatoeba 和 JW300）提高了 4.83 分 [34]，在 597 个样本的评估集上，大约达到 70 % 的语义翻译准确性。这些结果强调了预训练变压器模型在增强低资源 NMT 系统中的潜力。

局限性与未来工作。本研究中遇到的主要限制是 GPU 内存的限制，这限制了我们基线模型和迁移学习模型的训练深度。获得更高容量的计算资源可能会进一步提升模型的收敛性和性能。

对于未来的工作，我们提出了几个方向：

- 增加词汇量并扩展架构以支持多种低资源语言的多语言神经机器翻译。
- 探索更高级的注意力机制，例如局部注意力 [46]，这可能为长句提供更好的对齐。
- 研究其他解码策略，如最小贝叶斯风险（MBR）和基于 ROUGE 的解码，这些策略可能比束搜索（beam search）表现更好，同时计算负担较小。
- 使用图神经网络（GNNs）集成句法信息，这在编码语言结构和提升语义翻译质量方面显示出潜力 [4, 75]。

总之，我们的工作为英语-伊博语翻译建立了一个强有力的基准，并为在低资源环境中构建可扩展、准确和资源高效的神经机器翻译系统提供了实用指导。

ACKNOWLEDGMENTS

This research was conducted during the author's MSc program at the Moscow Institute of Physics and Technology (MIPT), Russia, in July 2022. The author gratefully acknowledges MIPT for providing the funding and computational resources that supported this work. Special thanks to Dr. Biswarup Das for his dedicated supervision, guidance, and mentorship throughout the research and the entire master's program.

REFERENCES

[1] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. 2018. Probabilistic fasttext for multi-sense word embeddings. arXiv preprint arXiv:1806.02901 (2018).
[2] Priya Ba, JM Nandhini, and Gnanasekaran Tc. 2021. An Analysis of the Applications of Natural Language Processing in Various Sectors. (2021).
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014).
[4] Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. arXiv preprint arXiv:1704.04675 (2017).
[5] Randall D Beer. 1997. The dynamics of adaptive behavior: A research program. Robotics and Autonomous Systems 20, 2-4 (1997), 257–289.
[6] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. Advances in Neural Information Processing Systems 13 (2000).
[7] Siddharth Bhatia, Bryan Hooi, Minji Yoon, Kijung Shin, and Christos Faloutsos. 2020. Midas: Microcluster-based detector of anomalies in edge streams. In Proceedings of the AAAI conference on artificial intelligence , Vol. 34. 3242–3249.
[8] Lukas Biewald. 2020. Experiment Tracking with Weights and Biases. https://www.wandb.com/ Software available from wandb.com.
[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics 5 (2017), 135–146.
[10] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2016. Quasi-recurrent neural networks. arXiv preprint arXiv:1611.01576 (2016).
[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.
[12] Jason Brownlee. 2017. Why One-Hot Encode Data in Machine Learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. Accessed: 2022-03-28.
[13] Yen-Yu Chang, Pan Li, Rok Sosic, MH Afifi, Marco Schweighauser, and Jure Leskovec. 2021. F-fade: Frequency factorization for anomaly detection in edge streams. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining . 589–597.
[14] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. arXiv preprint arXiv:1804.09849 (2018).
[15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).
[16] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning . 160–167.
[17] Mario Costa, Eros Pasero, Federico Piglione, and Daniela Radasanu. 1999. Short term load forecasting using a synchronously operated recurrent neural network. In IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339) , Vol. 5. IEEE, 3478–3482.

[18] Marta R Costa-Jussa and José AR Fonollosa. 2015. Latest trends in hybrid machine translation and its applications. Computer Speech & Language 32, 1 (2015), 3–10.
[19] Pradeep K. Debabala S., Prasant K. 2019. Advances in Intelligent Systems and Computing . O'Reilly Media, Inc.
[20] Deeplearing.AI. 2021. NLP Specialization. https://www.deeplearning.ai/. Accessed: 2022-04-20.
[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
[22] Ocheme Anthony Ekle and William Eberle. 2024. Anomaly Detection in Dynamic Graphs: A Comprehensive Survey. ACM Transactions on Knowledge Discovery from Data (2024).
[23] Ocheme Anthony Ekle and William Eberle. 2024. Dynamic PageRank with Decay: A Modified Approach for Node Anomaly Detection in Evolving Graph Streams. In The International FLAIRS Conference Proceedings , Vol. 37.
[24] Ocheme Anthony Ekle, William Eberle, and Jared Christopher. 2025. Adaptive DecayRank: Real-Time Anomaly Detection in Dynamic Graphs with Bayesian PageRank Updates. Applied Sciences 15, 6 (2025), 3360.
[25] Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512 (2019).
[26] Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. arXiv preprint arXiv:2004.00648 (2020).
[27] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In International Conference on Machine Learning . PMLR, 1243–1252.
[28] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. Journal of machine learning research 3, Aug (2002), 115–143.
[29] C Lee Giles, Steve Lawrence, and Ah Chung Tsoi. 1997. Rule inference for financial prediction using recurrent neural networks. In Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr) . IEEE, 253–259.
[30] Palash Goyal, Sumit Pandey, and Karan Jain. 2018. Deep learning for natural language processing. New York: Apress (2018).
[31] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning . 369–376.
[32] Philip A Gregory, Andrew G Bert, Emily L Paterson, Simon C Barry, Anna Tsykin, Gelareh Farshid, Mathew A Vadas, Yeesim Khew-Goodall, and Gregory J Goodall. 2008. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. Nature cell biology 10, 5 (2008), 593–601.
[33] Joel C Heck and Fathi M Salem. 2017. Simplified minimal gated unit variations for recurrent neural networks. In 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS) . IEEE, 1593–1596.
[34] Helsinki-NLP. 2021. OPUS-MT for English-Igbo Translation. https://huggingface.co/Helsinki-NLP/opus-mt-en-ig. Accessed: 2025-04-23.
[35] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.
[36] John Hutchins. 1997. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. a chronology. Machine Translation 12, 3 (1997), 195–252.
[37] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. arXiv preprint arXiv:1610.10099 (2016).
[38] Shubham Khandelwal, Benjamin Lecouteux, and Laurent Besacier. 2016. Comparing GRU and LSTM for automatic speech recognition . Ph. D. Dissertation. LIG.
[39] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
[40] E Kumar. 2011. Natural Language Processing; IK International Pvt. Ltd.: New Delhi, India (2011).
[41] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. Advances in neural information processing systems 29 (2016).
[42] Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. 2019. Application of long short-term memory (LSTM) neural network for flood forecasting. Water 11, 7 (2019), 1387.
[43] Robert B Lees. 1957. Syntactic structures.
[44] ED Liddy. 2001. Natural language processing in Encyclopedia of Library and Information Science 2nd ed., New York: Marcel Decker.
[45] Yangshengyan Liu, Fu Gu, Xinjian Gu, Yijie Wu, Jianfeng Guo, and Jin Zhang. 2022. Resource recommendation based on industrial knowledge graph in low-resource conditions. International Journal of Computational Intelligence Systems 15, 1 (2022), 42.
[46] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015).
[47] Larry Medsker and Lakhmi C Jain. 1999. Recurrent neural networks: design and applications . CRC press.
[48] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies . 746–751.
[49] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. Journal of the American Medical Informatics Association 18, 5 (2011), 544–551.

[50] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models. arXiv preprint arXiv:2201.08471 (2022).
[51] Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. arXiv preprint arXiv:1703.01619 (2017).
[52] Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. arXiv preprint arXiv:1708.09803 (2017).
[53] MD Okpor. 2014. Machine translation approaches: issues and challenges. International Journal of Computer Science Issues (IJCSI) 11, 5 (2014), 159.
[54] 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 of the Association for Computational Linguistics . 311–318.
[55] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) . 1532–1543.
[56] Duc Truong Pham and Dervis Karaboga. 1999. Training Elman and Jordan networks for system identification using genetic algorithms. Artificial Intelligence in Engineering 13, 2 (1999), 107–117.
[57] Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. arXiv preprint arXiv:2405.12819 (2024).
[58] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21, 1 (2020), 5485–5551.
[59] Thilina Rajapakse. 2020. Simple Transformers. https://simpletransformers.ai/docs/seq2seq-model/. Accessed: 2022-03-28.
[60] Nils Reimers and Iryna Gurevych. 2020. Multilingual sentence embeddings using distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) . 4512–4525.
[61] Roger C Schank and Larry Tesler. 1969. A conceptual dependency parser for natural language. In International Conference on Computational Linguistics COLING 1969: Preprint No. 2 .
[62] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. Neural networks 61 (2015), 85–117.
[63] Lane Schwartz. 2018. The history and promise of machine translation. Innovation and expansion in translation process research (2018), 161.
[64] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018).
[65] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems 27 (2014).
[66] HuggingFace Tatoaba. 2022. Eng-Fra Benchmark. https://huggingface.co/Helsinki-NLP/opus-mt-en-fr. Accessed: 2022-04-25.
[67] Tolga Uslu, Alexander Mehler, Daniel Baumartz, and Wahed Hemati. 2018. fastsense: An efficient word sense disambiguation classifier. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) .
[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).
[69] Lilapati Waikhom and Ripon Patgiri. 2021. Graph neural networks: Methods, applications, and opportunities. arXiv preprint arXiv:2108.10733 (2021).
[70] Sheng Wang, Dian Yu, Yizhong Zhang, Yejin Choi, Jianfeng Gao, et al. 2023. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv preprint arXiv:2302.13952 (2023).
[71] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019).
[72] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Paul Barham, Xiaodong Guo, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934 (2021).
[73] Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics . 523–530.
[74] Baosong Yang, Longyue Wang, Derek Wong, Lidia S Chao, and Zhaopeng Tu. 2019. Convolutional self-attention networks. arXiv preprint arXiv:1904.03107 (2019).
[75] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. arXiv preprint arXiv:2007.08742 (2020).
[76] Hongwei Zhao, Zhongxin Chen, Hao Jiang, Wenlong Jing, Liang Sun, and Min Feng. 2019. Evaluation of three deep learning models for early crop classification using sentinel-1A imagery time series—A case study in Zhanjiang, China. Remote Sensing 11, 22 (2019), 2673.
[77] Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou. 2016. Minimal gated unit for recurrent neural networks. International Journal of Automation and Computing 13, 3 (2016), 226–234.
[78] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201 (2016).