ChatGPT: Jack of all trades, master of none

摘要

原复现论文是关于 ChatGPT 在多任务自然语言处理中的评估，研究者使用不同的提示生成方式来引导 ChatGPT 进行情感分析、情绪检测和词义消歧等任务。通过后处理 ChatGPT 输出并与基准模型进行比较，评估了其在多任务处理中的表现。结合自己的研究方向，我重点关注其中词义消歧任务进行评估和拓展。对该任务的数据、基准和流程进行了更深入的了解。通过该研究，旨在纵向学习和优化大型语言模型在 WSD 任务上的应用，并探索其作为新的基准的潜力。

关键词： ChatGPT；词义消歧；提示工程

1 引言

OpenAI 发布了聊天生成预训练 Transformer(ChatGPT)，彻底改变了人工智能与人机交互的方法。与聊天机器人的第一次接触揭示了它在各个领域提供详细而准确的答案的能力。一些关于 ChatGPT 评估的出版物测试了其在众所周知的自然语言处理 (NLP) 任务上的有效性。然而，现有的研究大多是非自动化的，并且测试规模非常有限。这篇复现文章的主要内容是关于使用 ChatGPT 进行多任务自然语言处理的评估。研究者使用了不同的提示生成模式来指导 ChatGPT 回答特定的自然语言处理任务，如情感分析、情绪检测、词义消歧等。他们通过对 ChatGPT 的输出进行后处理，将其转化为预定义的标签，并与基准模型进行比较和评估。研究者还介绍了他们的实验设置和使用的 API 工具。总体而言，这篇文章旨在评估 ChatGPT 在多任务处理中的表现。

结合我自己的研究方向，我将重点对其中关于 ChatGPT 对于词义消歧（WSD）任务的评估进行复现和纵向的评估改进。其中词义消歧 (WSD) 是文本语义分析的一项基本任务。它旨在从特定上下文中出现的目标词的预定义候选词义集中识别出正确的词义。由于其在多个下游 NLP 任务或网络文本挖掘（例如机器翻译和信息提取）中的应用，近年来获得了关注的一项重要任务 [1]。

预训练语言模型目前已经在自然语言处理领域得到广泛应用，但其应用范围仍有不少未知空间。原文探讨了将预训练语言模型应用于各种 NLP 任务的可能性，是大语言模型对该 NLP 领域的一次尝试和探索。我自己的研究方向是关于词义消歧（WSD）的评估，对这方面的数据、基准及任务流程都比较熟悉，通过复现和改进这篇文章在 WSD 任务上的评估，我可以纵向的学习使用 ChatGPT 等大语言模型对本领域的任务进行优化，甚至作为新的基准。

2 相关工作

2.1 传统词义消歧方法

词义消歧（WSD）是文本语义分析的基本任务。它旨在从一个特定上下文中的预定候选词义集合中识别正确的词义。这是一个重要的任务，近年来在下游 NLP 任务或网页文本挖掘中得到了广泛应用，例如机器翻译和信息提取 [1]。

目前已经提出的几种具有潜力和希望的 WSD 方法可以主要分为监督方法和基于知识的方法。监督方法利用带有词义标记的语料库，并学习在语料库中的歧义词上执行词义分类。虽然这些方法在实验中显示出有希望的结果 [2]，但当目标词义在训练集中缺失时，性能会显著下降。

基于知识的方法利用现有的词汇资源，如 WordNet [3]。词汇资源可以提供两种类型的信息：文本信息（例如，词义释义）和结构信息（例如，上位词和下位词）。尤其是词义释义提供了一个词义的简短说明。在 WSD 中词义释义的使用具有很长的历史，可以追溯到 Lesk 的工作 [4]。在此工作之后，一系列的研究已经结合使用词义释义来完成 WSD 任务 [5]，并且经验证实使用词义释义信息的好处。

WSD 中的一个较新的趋势是将词汇资源的信息与监督学习方法相结合。词汇资源（例如，词义释义）与监督学习方法特别是与神经网络相结合，已经在 WSD 中有很好的表现提升。在这些工作中，Luo 等人 [6] 使用全局向量（GloVe）[7] 作为词嵌入。同时，基于 Transformer(BERT) [8] 以及其拓展模型的双向编码器表征通过多头注意力机制可以进行更深层次的上下文建模，并在最新的 WSD 模型中具有很好的表现。

最近的研究通过使用标记语料库和词汇资源信息来获得更好的性能。BEM [9] 首次在 WSD 中应用两个转换器分别对上下文和简介进行编码。基于 BEM，SACE [10] 通过考虑上下文中邻近词的词义来增强简介表示。ESCHER [11] 将 WSD 重新构造为一个跨度提取问题。ConSec [12] 通过引入目标词的上下文来进一步修改 ESCHER 的任务设置。Su [13] 提出了一种 Z 加权策略来缓解数据不平衡。Zhang [14] 提出了一种具有独立约束机制的分解表示方法。KELESC [15] 通过从 WordNet 中获取附加语义知识来丰富目标词的上下文，并使用基于局部自注意力变换器的局部自注意力变换器对其进行编码。

2.2 Transformer 到 ChatGPT 的发展变革

近年来，Transformer 类型的模型架构主导了自然语言处理（NLP）领域 [16–18]。在此之前，循环神经网络（如 LSTM）被用于解决各种现有的 NLP 问题 [19,20]。循环神经模型无法捕捉数据序列中的远程依赖关系，例如文本开头或结尾处出现的信息 [21]。此外，它们的架构不允许有效地并行化训练和推理过程 [22]。对上述问题新的解决方案正是 Transformer 架构，在最初作为序列到序列任务的编码器-解码器模型中提出 [16]。这样一个模型具有使用注意机制捕捉文本中远程关系以及通过矩阵运算轻松并行计算的优势。随着更强大的 GPU 和 TPU 设备被开发出来 [23]，就可以创建参数越来越多、在越来越多任务上达到人类性能水平的模型了。然而，最显著质量改进是通过在从互联网获取到大量文本上进行无监督预训练语言模型实现的。在基于 BERT 的模型中，预训练任务包括预测掩盖标记和后续句子 [24]。在自回归模型中，预训练任务已被改为预测下一个词，这样可以屏蔽注意力层，使模型仅基于过去

的值来预测未来的值 [25]。

生成式预训练（GPT [26]）是基于 Transformer 架构的最早的自回归生成模型之一。从原始的 Transformer 中，GPT 只使用了解码器堆栈，并将双向自注意力转换为单向。这样的模型可以执行基于生成新文本的所有任务，例如翻译、摘要或回答问题。在 GPT-2 中，对该概念进行了扩展，进行了几项技术改进，消除了将模型微调到下游任务时存在的可传递性问题，并引入了多任务训练 [27]。此外，输入上下文长度从 512 增加到 1024，预训练数据增加到 40GB，但模型参数总数从 117M（GPT）飙升至 1.5B（GPT-2）。结果表明，GPT-2 显示出在不需要大规模监督训练数据的情况下能够解决许多新任务的能力。GPT-3 模型主要有两个因素区别于前者：模型参数数量增加至 175B，并且使用 45TB 文本数据进行预训练。该模型在零样本和少样本场景中表现出色 [28]。

进一步实现将模型响应与人类需求匹配的一步是创建 InstructGPT 模型 [29]。其主要创新集中在替代模型微调方法上，特别是通过人类反馈进行强化学习（RLHF）。该解决方案使用人类反馈作为更新模型参数的奖励信号。OpenAI 聘请了 40 名具有敏感言论标记、排名模型答案质量、敏感示范写作以及能够识别不同群体敏感言论等高度一致性水平评注员。他们的任务是描述针对不同提示期望得到何种回答，并根据给定提示对系统生成多个回复进行排序以训练奖励模型。在第三步中，使用近端策略优化（PPO）进行强化学习以进一步提高模型质量。结果显示，用户对 InstructGPT 的响应比 GPT-3 更为青睐。这项工作的一个结论是，在公开可用的 NLP 基准数据集上，模型质量较 SOTA 模型差。然而，InstructGPT 的作者发现基准 NLP 任务并不能反映大多数人对语言模型的真正期望 [29]。只有 18% 使用 OpenAI API 的用户查询了与典型 NLP 任务相熟悉的 GPT-3 模型，其中大部分是分析性任务。另一方面，在评估 InstructGPT 时仅使用了少部分流行的 NLP 数据集 [29]。

InstructGPT 的最新版本之一是 ChatGPT 模型（图 1），它很可能利用更多用户反馈来处理更广泛种类的任务。目前对该模型的构建了解甚少，但其系统的出色质量使其广受欢迎。有趣的是，在 InstructGPT 中，基础模型只有 35B 参数 [29]。然而，在对话任务中，它提供比拥有 175B 参数的 GPT3 模型更好的答案。这表明从人类那里收集数据进行监督模型微调具有很高的相关性 [29]。ChatGPT 的下一代 GPT-4 [30] 将是一个更大的模型，除了文本输入外还可以接收图像作为输入。

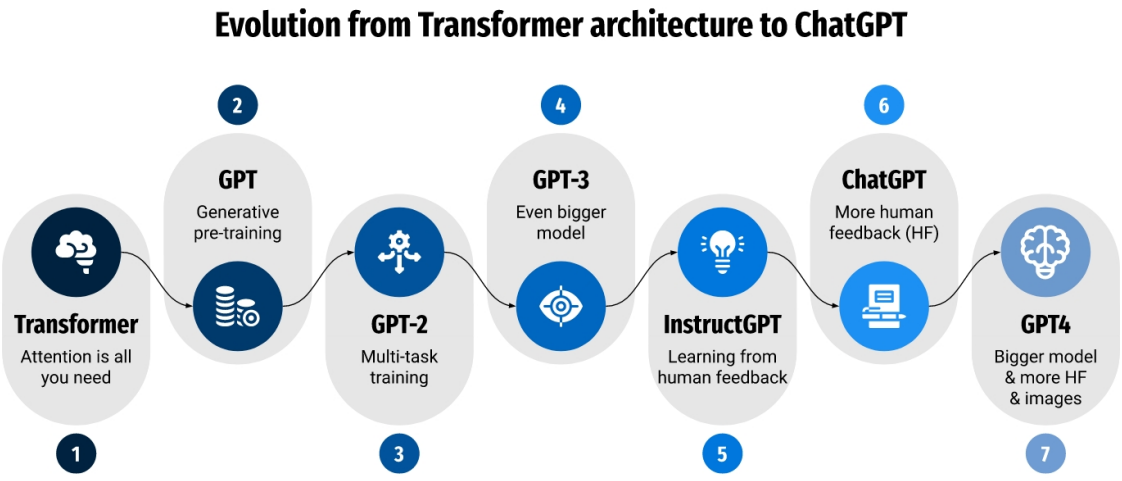

图 1. 基于 Transformer 架构的自回归模型的发展

3 数据与方法

3.1 数据框架概述

目前 WSD 领域内对英文的全词消歧所用到的数据集评估框架大多为 Raganato 等人 [31] 提出的框架 WSD Evaluation Framework，这篇复现的论文使用的也是它。WSD Evaluation Framework 是用于评估词义消歧系统性能的一个框架，其设计旨在为词义消歧系统性能评估提供一种统一的方法。它可以通过使用已知语料库（如 SemCor）来训练词义消歧系统，并使用不同的评估数据集（例如 SemEval-2007、Senseval-2、Senseval-3、SemEval-2013 和 SemEval-2015，这些数据集的全集为 ALL）来对系统进行测试和评估。这个框架允许研究人员比较不同的词义消歧系统在不同数据集上的性能表现，以便更好地理解和解释系统的优劣势。通过使用 WSD Evaluation Framework，研究人员可以评估他们的系统在真实世界语境下的表现，并且可以获得系统在每个数据集以及所有数据集上的 F1 分数，从而进行全面的性能分析和比较。

但是由于很多之前的工作已经在这个数据上的表现达到了人类标注者所能达到的最好表现。因此，我的复现工作不仅复现了 ChatGPT 在这个评估框架上的评估结果，同时引入了一个新的由 Sapienza NLP Group [32] 提出的更有挑战性的数据标准 wsd-hard-benchmark 上。

wsd-hard-benchmark 包含 42D、ALL_NEW、S10_NEW、softEN 和 hardEN 五个子数据集，其中 42D 的源文本样本被抽取以代表不同文本领域，具体来说就是 BabelNet 4.0 [33] 定义好的 42 种领域，可以当作一个跨领域的评估数据集。ALL_NEW 和 S10_NEW 是来自之前原始数据集的修订版本。hardEN 是之前所有评估模型在 ALL 上的预测错误实例的集合，而 softEN 则是剩下的由某些系统正确预测的实例组成。

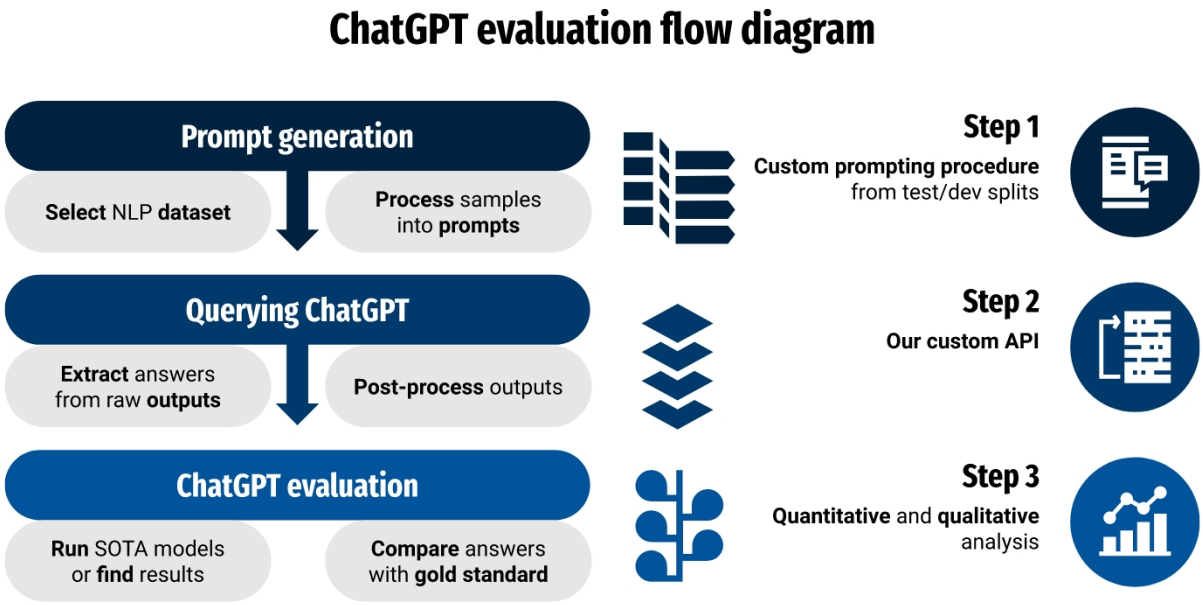

图 2. ChatGPT 评估流程图

3.2 方法概述

如图 2所示，为针对 NLP（包含 WSD）任务在 ChatGPT 的基础上进行提示工程的整个工作流程。特别地，我们针对词义消歧任务，可以把它看成一个多分类选择问题，对于给定

的包含目标词的上下文和目标词的候选词义，我们只需让 ChatGPT 来选出其中最符合上下文语境的正确词义即可。具体而言，在选定评估的测试集后，将数据集中说要得到的上下文和目标词以及候选词义通过一系列方式提取出来，然后据此生成提示词，输入给 ChatGPT，得到回复后，对回复得到的输出结果提取出由 ChatGPT 选出的正确词义，然后进行处理并评估，最后，与现有的 SOTA 模型做对比。

Chat 49. Task: WSD. Case 3.

Prompt

Which meaning of the word "peculiar" is expressed in the following context: *The art of change-ringing is peculiar to the English, and, like most English peculiarities, unintelligible to the rest of the world. Dorothy L. Sayers, " The Nine Tailors" ASLACTON, England– Of all scenes that evoke rural England, this is one of the loveliest: An ancient stone church stands amid the fields, the sound of bells cascading from its tower, calling the faithful to evensong. The parishioners of St. Michael and All Angels stop to chat at the church door, as members here always have.* The meanings are as follows:

- 'peculiar%5:00:00:characteristic:00': characteristic of one only; distinctive or special;
- 'peculiar%5:00:00:strange:00': beyond or deviating from the usual or expected;
- 'peculiar%5:00:00:unusual:00': markedly different from the usual.

Return only the key of the most relevant meaning.

ChatGPT answer

peculiar%5:00:00:specific:00

Expected answer

peculiar%5:00:00:specific:00

图 3. WSD 提示工程示例

如图 3所示，为原文中最原始的一种向 ChatGPT 提问的方法。其中我们需要消歧的目标词为 "peculiar"，它包含在上下文中，从 WordNet 中可以提取出其有三种词义，在通过 Prompt 向 ChatGPT 提问后，ChatGPT 最终返回在这个上下文中最相关的词义候选项的 key 为 "peculiar%5:00:00:specific:00"，与期望返回的正确词义一致，代表预测正确。

4 复现细节

4.1 与已有开源代码对比

原论文中开放了 notebook 版本的源代码，但我们并没有采取它的源代码，而是在我使用了我之前工作中和是写的自己的代码来完成的这个任务，只是借鉴了它源码中原始 prompt 的问法。同时，我的代码中还参考了由 wsd-hard-benchmark 作者提出评估方法的代码，计算了最后在各种测试集上评估的宏平均 F1 分数和微平均 F1 分数作为评价指标，因为这篇文章提出了使用宏平均 F1 分数作为微平均 F1 分数的替代方法，以更好地考虑 WSD 评估中最不常见的词义。

4.2 改进与创新

首先，原文只在 Raganato 等人提出的最基础的基准 [31] 上做了词义消歧的评估，我将会将其延伸到更具挑战性的词义消歧评估框架 wsd-hard-benchmark [32] 上评估，并与最新的传统 SOTA 模型做对比。

另外，针对 wsd-hard-benchmark 中最典型的两个数据集 42D 和 hardEN，我通过提示工程，并利用一些其他信息（如上下文跨度、翻译信息等）和方法（如思维链等）对原始的原文中最原始的 prompt 进行了改进并得了更好的评估结果。如图 4所示，为我改进的 4 种提示工程策略。其中每一方法都对原文中最原始的 prompt 都做了单一变量式的改进。

Methods	Prompt Engineering
original	Which meaning of the word "{word}" between {start_tag} and {end_tag} in the following context is expressed: "{sentence}" The meanings are as follows: {senses}. Return only the key of the most relevant meaning.
Method-1 (角色指定+情感激励)	{"role": "system", "content": "You are a English linguist."}, {"role": "user", "content":"I have a word sense disambiguation task for you, which is very important for my career. <original>"}
Method-2 (直接选 gloss 而非 key)	Which meaning of the word "{word}" between {start_tag} and {end_tag} in the following context is expressed: "{sentence}" The meanings are as follows: {ordered_senses}. Make sure to return the complete option of the most relevant meaning.
Method-3 (思维链利用跨语言词级翻译信息)	<original> The following methods may help you choose the correct meaning: First, rate the semantic relatedness of each meaning to the given context. If there are meanings that are semantically close and difficult to distinguish, you can try translating the target word into other languages to help align and judge. Finally, select the meaning that best matches the target word and context.
Method-4 (拓展上下文)	Which meaning of the word "{word}" between {start_tag} and {end_tag} in the following context is expressed: "{prefix_sententce}" + "{target_sentence}" + "{suffix_sentence}" The meanings are as follows: {senses}. Return only the key of the most relevant meaning.

图 4. 改进的方法核心

5 实验结果分析

针对论文中得到的 WSD 在传统评估框架 WSD Evaluation Framework 中 ALL 上最终的宏平均 F1 分数为 73.0，我自己复现的后评估的结果为 73.25。这个结果和论文原本的结果相当，有误差的原因一是由于不知道论文作者用的 ChatGPT 的版本，我使用的最新接口与原文作者之间调用的 ChatGPT 的接口并不相同。另外的一个原因就是 ChatGPT 返回的结果也具有小部分的随机性，会在个数位及后面更小的位数产生影响。

针对我的改进的第一个部分，测试 ChatGPT 在新的有挑战性的数据集框架 wsd-hard-benchmark 上的结果如图 5所示，其中 RTWE 为当前由 zhang [1] 等人提出的最新的传统 SOTA 模型。从中可以看出，在 ALL_NEW、S10_NEW 和 softEN 这些数据集上的表现 ChatGPT 不如传统模型，这和论文作者自己的观点也是一致的，他认为 ChatGPT 虽然知识面很广，但是做不到很精。但出乎意料的的是在 42D 和 hardEN 上的表现却大幅超过了传统 SOTA 模型。达成这个表现可能的原因是，由于 42D 和 hardEN 里面的目标词其实有些还是比较常见的，但比较麻烦的是有些词义可能比较相近或者有一些误导性，同时有些词义数量很多，达十几条，长度也比较长，造成了选择的困难。一个直观的解释是大模型因为是经过海量语料库预训练，它本身就对一些常见单词的通用词义已经具备很好的消歧和理解能力，然后泛化能力也很强，所以导致表现效果没有那么差。

dataset	#inst	RTWE (base)		RTWE (large)		ChatGPT (gpt-3.5-turbo-1106)	
		M-F1	m-F1	M-F1	m-F1	M-F1	m-F1
ALL_NEW	4,917	77.89	81.55	80.45	83.32	71.3	73.52
S10_NEW	955	73..82	79.16	75.45	81.05	74.46	74.46
42D	370	48.99	45.95	52.93	49.73	63.92	62.16
softEN	5,766	80.6	84.91	83.2	86.78	75.33	76.83
hardEN	476	8.81	7.98	13.28	11.34	36.07	34.66

图 5. ChatGPT 在新数据集上和传统 SOTA 模型的比较

针对我改进的第二个部分，对原论文中的 prompt 进行改进，得到的各种方法的在 42D 和 hardEN 的表现效果如图 6所示。其中前两行中的数据分别为各个方法在数据集中测试后得到的宏平均 F1 分数/微平均 F1 分数，后两行得出各个方法测评的结果分数和原始方法的差异。我们可以看到特别是方法 2 最两个数据集的表现都有较明显的提升，而方法 2 指的是将 prompt 中通过限定目标词义项的 key 修改成直接获取它的词义本身。得到提升可能的原因是在 WordNet 中词的词义项表示形式的 key 过于复杂，对模型的判断产生了负面的效果。而为什么方法 1 对 42D 数据集有明显提升，需要进一步实验探索探究原因。

dataset	Methods				
	Origial	Method-1	Method-2	Method-3	Method-4
42D	63.92 / 62.16	70.04 / 68.38	69.58 / 67.30	67.67 / 65.95	65.95 / 63.51
hardEN	36.07 / 34.66	37.79 / 36.13	42.89 / 40.71	39.09 / 36.55	34.29 / 31.72
42D	0 / 0	**6.12 / 6.22**	**5.66 / 5.14**	3. 75 / 3.79	2.03 / 1.35
hardEN	0 / 0	1.72 / 1.47	**6.82 / 6.05**	3.02 / 1.89	-1.78 / -3.59

图 6. ChatGPT 在新数据集上和传统 SOTA 模型的比较

6 总结与展望

本报告对选定的目标文章的复现和改进进行了书面化的梳理，其中最核心的内容其实是使用 ChatGPT 对当前词义消歧领域的传统数据集及新的挑战性数据集做了相对完整的评估和改进。希望自己通过复现和改进这篇文章在 WSD 任务上的评估，可以纵向的学习使用 ChatGPT 等大语言模型对本领域的任务进行优化，甚至作为新的基准以备后续研究。

另外，本文的评估中也存在一些不足之处待改进和日后完善。一是 ChatGPT 在评估过程中会有一定的随机性来影响最后的结果，后面会想办法消除这种影响。二是通过提示工程提示在测试集上的表现还有更多的方法去尝试，如利用目标词附件的显示词义信息、从文档的角度匹配相似度高的相关句子来丰富语义特征、通过 ChatGPT 对目标文段的进行摘要或是主题概括来帮助其聚焦等，这些方法都有待去尝试和实践来改进目前效果。

参考文献

[1] Xuefeng Zhang, Richong Zhang, Xiaoyang Li, Fanshuang Kong, Junfan Chen, Samuel Mensah, and Yongyi Mao. Word sense disambiguation by refining target word embedding. In *Proceedings of the ACM Web Conference 2023*, pages 1405–1414, 2023.

[2] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Jan 2017.

[3] George A. Miller. Wordnet. In *Proceedings of the workshop on Speech and Natural Language - HLT ' 91*, Jan 1992.

[4] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th annual international conference on Systems documentation - SIGDOC ' 86*, Jan 1986.

[5] Satanjeev Banerjee and Ted Pedersen. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*, page 136–145. Jan 2002.

[6] Fuli Luo, Tian-Yu Liu, Xia Qin, Baobao Chang, and Zhifang Sui. Incorporating glosses into neural word sense disambiguation. *Cornell University - arXiv,Cornell University - arXiv*, May 2018.

[7] Jeffrey Pennington, Richard Socher, and ChristopherD Manning. Glove: Global vectors for word representation.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[9] Terra Blevins and Luke Zettlemoyer. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jan 2020.

[10] Ming Wang and Yinglin Wang. 2021-acl-sace-word sense disambiguation- towards interactive context exploitation from both word and sense perspectives. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Jan 2021.

[11] Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021-acl-esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jan 2021.

[12] Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021-emnlp-consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan 2021.

[13] Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. Rare and zero-shot word sense disambiguation using z-reweighting.

[14] Junwei Zhang, Ruifang He, Fengyu Guo, Jinsong Ma, and Mengnan Xiao. Disentangled representation for long-tail senses of word sense disambiguation. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, Oct 2022.

[15] Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*, Jun 2017.

[17] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: a systematic survey. *Artificial Intelligence Review*, Aug 2022.

[18] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, page 111–132, Jan 2022.

[19] Rie Johnson and Tong Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2016.

[20] Basemah Alshemali and Jugal Kalita. Improving the reliability of deep neural networks in nlp: A review. *Knowledge-Based Systems*, page 105210, Mar 2020.

[21] Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, page 325–338, Apr 2019.

[22] ZacharyC. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv: Learning,arXiv: Learning*, May 2015.

[23] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183, 2020.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] PeterJ. Liu, M.A. Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *International Conference on Learning Representations,International Conference on Learning Representations*, Jan 2018.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[30] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[31] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April 2017. Association for Computational Linguistics.

[32] Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. Nibbling at the hard core of Word Sense Disambiguation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[33] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, page 217–250, Dec 2012.