中图分类号：　　单位代号：10280
密　　级：　　学　　号：20721213

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	语言模型驱动的文本隐写技术研究

作　　者　郑晓燕
学科专业　通信与信息系统
导　　师　吴汉舟
完成日期　2023 年 5 月

姓　　名：郑晓燕　　学号：20721213

论文题目：语言模型驱动的文本隐写技术研究

上海大学

本论文经答辩委员会全体委员审查,确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导　　师：

答辩日期：

姓　　名：郑晓燕　　学号：20721213

论文题目：语言模型驱动的文本隐写技术研究

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____日 期：_____

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密的论文在解密后应遵守此规定**）

签 名：_____导师签名：_____日期：_____

上海大学工学硕士学位论文

语言模型驱动的文本隐写技术研究

姓　　名：郑晓燕
导　　师：吴汉舟
学科专业：通信与信息系统

上海大学通信与信息工程学院

二 O 二三年五月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Research on Text Steganography Driven by Language Model

Candidate：Xiaoyan Zheng

Supervisor：Hanzhou Wu

Major：Communication and Information System

School of Communication and Information Engineering

Shanghai University

May, 2023

摘　要

过去的二十年见证了在线社交网络服务的快速发展，在社交网络中人们借助媒体实时分享日常生活的感受。这给隐蔽通信带来了极大的便利，隐蔽通信旨在可靠地将秘密信息传送给数据接收者而不引起监视者的怀疑。文本是日常生活中信息交换的重要媒介之一，然而由于其本身的低冗余性，使得文本隐写的研究更具挑战性，同时也更富有意义。与传统的方法相比，利用语言模型实现文本隐写不仅能够增加载密文本的嵌入容量，而且能够提高其安全性。在此背景下，本文研究语言模型驱动的文本隐写，相关的研究工作与成果如下：

(1) 针对主流的非自回归文本隐写方法的局限性，本文提出了一种基于BERT和一致性编码的自回归文本隐写算法，它在嵌入容量和系统安全性之间实现了更好的权衡。在提出的工作中，基于掩码语言模型，针对给定的文本使用一致性编码来弥补块编码的缺点，这样就可以编码任意大小的候选词集，并利用概率分布进行信息隐藏。针对需要嵌入信息的掩码位置，以自回归的方式预先进行单词填充，以增强上下文之间的联系，从而保证文本质量。实验结果表明，与非自回归文本隐写方法相比，该工作在保证安全性的同时提高了载密文本的流畅性，并在一定程度上增加了嵌入容量。

(2) 虽然现有文本信息隐藏方法能够完美地提取秘密信息，但载体文本会出现永久性失真，针对上述问题，本文提出了一种基于掩码语言模型的可逆文本信息隐藏算法，该算法可以将嵌入的信息和载体文本从载密文本中完美地提取出来。主要思想是数据隐藏者使用掩码语言模型生成载密文本，数据接收者通过收集某些位置的单词来重建载体文本，并对其他位置的词进行同样的操作以提取秘密信息。实验结果表明，载体文本和秘密信息可以被成功嵌入和提取。同时，携带秘密信息的载密文本具有良好的流畅性和语义质量，并且达到了较好的安全性。此外，数据隐藏者和数据接收者之间不需要共享语言模型，这减少了双方共享的边信息，因此该方法具有良好的应用前景。

关键词：隐蔽通信，文本隐写，语言模型，可逆信息隐藏

ABSTRACT

The past two decades have witnessed the rapid development of online social networking services, where people use media to share their daily life feelings in real time. This brings great convenience to covert communications, which are designed to reliably transmit secret information to data recipients without arousing the suspicion of watchers. Text is one of the important media for information exchange in daily life. However, due to its low redundancy, the study of text steganography is more challenging and meaningful. Compared with traditional methods, using language model to implement text steganography can not only increase the payload of ciphertext, but also improve its security. In this context, this dissertation studies language model-driven text steganography. The related research work and achievements are as follows:

(1) Aiming at the limitations of the mainstream non-autoregressive text steganography methods, this dissertation proposes an autoregressive text steganography algorithm based on BERT and consistency coding, which achieves a better trade-off between embedding payload and system security. In this dissertation, based on the introduction of the masked language model, given a text, we use consistency coding to make up for the shortcomings of block coding used in the previous work so that we can encode arbitrary-size candidate token set and take advantage of the probability distribution for information hiding. The masked positions to be embedded are filled with tokens determined by an autoregressive manner to enhance the connection between contexts and therefore maintain the quality of the text. Experimental results have shown that compared with related works, the proposed work improves the fluency of the steganographic text while guaranteeing security and also increases the embedding payload to a certain extent.

(2) Although the existing text data hiding methods can perfectly extract the secret information, the cover text will be permanently distorted. In view of the above problems, this dissertation proposes a reversible data hiding algorithm in text based on the masked

language model, which can convert the embedded information and the original cover text are perfectly retrieved from the steganographic text. The main idea of the proposed method is to use a masked language model to generate such a marked text that the cover text can be reconstructed by collecting the words of some positions and the words of the other positions can be processed to extract the secret information. Our results show that the original cover text and the secret information can be successfully embedded and extracted. Meanwhile, the marked text carrying secret information has good fluency and semantic quality, indicating that the proposed method has satisfactory security, which has been verified by experimental results. Furthermore, there is no need for the data hider and data receiver to share the language model, which significantly reduces the side information and thus has good potential in applications.

Keywords: Covert Communication, Text Steganography, Language Model, Reversible Data Hiding

目　录

第一章　绪论

1.1 课题来源

本课题来源于国家自然科学基金青年项目“社交网络多用户协同的行为隐写”(项目编号：61902235)。

1.2 研究的背景与意义

随着信息技术的快速发展和广泛应用，人们享受便利的同时，也不可避免地面临着新的信息安全问题和潜在威胁。在数字通信网络中，由于其容易被恶意干扰和窃听，信息的安全性面临着较大的挑战。作为一种有效的秘密通信手段，信息隐藏运用数字媒体中的冗余部分来隐蔽地嵌入秘密信息。这样生成的载密媒体不会引入明显的伪影，因此，对媒体的使用不会产生实质性影响，同时能够实现秘密信息的传输和版权保护等多重目的[1]。

信息隐藏技术有两种常见的形式，分别是面向版权保护的数字水印和面向隐蔽通信的隐写。数字水印技术的功能在于将特殊的数字信号嵌入到多媒体中，以维护多媒体文件的真实性和完整性，并着重于版权声明和身份认证，以避免多媒体的非法复制。相较而言，隐写的主要目标是完成秘密数据的隐藏，以实现在感知上和统计上几乎达到难以察觉的效果，并起到保护数据安全性的作用[2, 3]。因此，隐写技术在军事、情报、金融等领域得到了广泛应用[4, 5]。

随着在线社交网络服务的快速发展，社交网络平台逐渐成为人们进行社交互动和共享生活体验的主要场所。这给隐蔽通信带来了极大的便利，隐蔽通信目的是可靠地将秘密信息传送给数据接收者而不引起第三方的怀疑。文本是不同语言和文化之间沟通的桥梁，在社交网络这样的公共场合，每个人依赖并利用文本来增长知识，交流沟通，展现思想。在这一场景下双方可以秘密地完成信息传输过程，数据接收者不需要进行令第三方怀疑的交互行为。这不仅具有更高的便捷性，而且能够保护情报人员的身份安全。

无论过去还是现在，文本是日常生活中信息交换的重要媒介之一，人们彼此

之间的交流和对话需要依靠文本来传达大量的信息。同时，自然语言在信息传输过程中具有较高的鲁棒性，即使在受到干扰的情况下，文本仍可以在公共通道中保持秘密传输而不失真。由上面两点可见，研究以文本作为隐藏信息的载体具有合理性和必要性。但是，由于文本本身的低冗余性，相比图像、音频和视频等数字媒体，文本隐写的研究更加复杂和具有挑战性[6-8]。

因此，在当前的信息安全领域，文本隐写具有不可替代的重要性。然而，要设计一种能够抵抗统计分析的文本隐写方法，更是一项极具挑战性的难题。为了有效解决这一问题并探索更高效、更安全的隐蔽通信系统框架，本文研究语言模型驱动的文本隐写具有一定的价值，从而更好地满足现代信息安全的需求。诸如此类的研究方向，将在未来的发展中扮演越来越重要的角色。

1.3 文本隐写国内外研究现状

隐写是实现信息隐藏的一种重要手段，受到了研究者们广泛的关注和重视。其常见的操作方式是将秘密数据隐藏到不同的数字载体中(如图像[9]、音频[10]、视频[11]和文本[12]等)。由于文本在我们日常生活中扮演着至关重要的角色，因此文本隐写对于隐写的发展具有巨大的推动作用。如图 1.1 所示，现阶段主流的文本隐写方法可以分为两类：即修改式文本隐写和生成式文本隐写。修改式文本隐写是通过对给定的自然文本进行轻微的修改来嵌入秘密信息，而生成式文本隐写则是直接生成带有秘密信息的新文本，这摆脱了原始文本的限制。

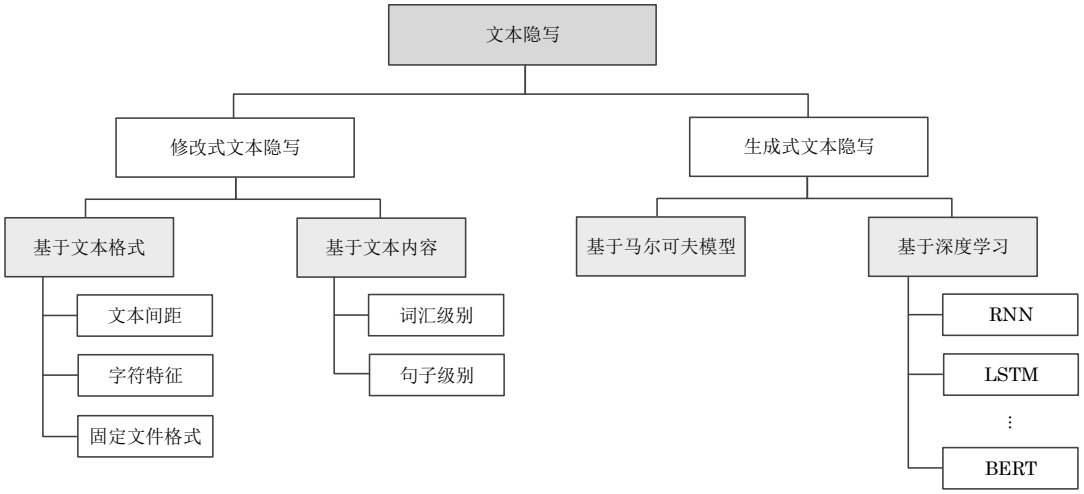


图1.1 文本隐写方法分类

1.3.1 修改式文本隐写

修改式文本隐写算法是目前最常见也是研究最充分的文本隐写策略，指的是用文本作为隐藏秘密信息的空间，对文本格式或者文本内容加以修改并嵌入秘密信息。修改式文本隐写方法可以根据修改的对象不同进一步分为两类，即基于文本格式的方法和基于文本内容的方法。

（1）基于文本格式的方法

基于格式的文本隐写通常通过调整文档的间距和字体样式等方式来隐藏秘密信息，从视觉上实现了信息的隐藏效果[13,14]。但是，它们在抗隐写分析方面存在一定的缺陷，对手可以通过重新排版文档、检测文本异常格式等技术手段来揭示秘密信息的存在。虽然基于格式修改的方法具有实现简单、易于操作等优点，但其抗隐写分析性较差，在保护信息安全方面需要进一步提高[15,16]。基于格式的文本隐写方法主要分为基于文本间距、基于字符特征和基于固定文件格式三类。

基于文本间距的隐写方法主要是通过微调文档中的行间距和字符间距来嵌入秘密信息。大量研究人员从改变文本格式的物理性质着手，提高了文本隐写的能力。例如，在 1995 年，Brassil 等人[17]提出了行移编码和字移编码，在不引起视觉注意的前提下，细微地移动行间距或字符间距达到隐藏信息的目的。行移编码是微调文本行与行之间的距离以嵌入秘密信息，字移编码则是在不影响文本整体阅读流畅性的前提下，嵌入特定字符并左右微调来传递秘密信息。在该算法中行移编码小于字移编码的嵌入容量，但行移编码的隐蔽性较好。

基于字符特征的文本隐写主要包括改变字体、字号和颜色等字符特征形式。Bhaya 等人[18]提出了一种在微软文档中修改文字字体的隐写方法，该方法工作原理是根据英语字体类型的相似性理论，用相似的字体替换原字体。该算法提升了信息隐藏容量，但受到字体特征的约束。Mahato 等人[19]提出了一种调整不可见字符的文本隐写方法。利用英文文本中存在大量的空格，在文本行中、段落中修改类似于空格的不可见字符进行嵌入，该方法得到的载密文本在视觉上易被发现，不可感知性较差。Tang 等人[20]由此想到，可以根据人的眼睛对蓝色最不敏感这一特性，提出了基于 RGB 的文本隐写方法，通过修改字符和下划线的 RGB 的 3 个通道最低位值来实现信息隐藏，该算法在视觉上具有良好的性能。

基于固定文件格式的隐写方法具有很强的局限性，可以利用不同类型文档，例如 Word，PPT（Powerpoint），PDF（Portable Document Format）等，自身特定的属性和功能实现信息的嵌入。针对微软 Word 的文档修订功能，Liu 等人[21]提出了一种利用变化跟踪技术在 Word 文档中进行数据隐藏的隐写方法。在一个正常的协作的场景中，该方法模拟一个谨慎的作者使用文档修订功能纠正错误。

针对 PPT 文档，注释页是 PPT 文档中每张幻灯片的重要组成部分，它为演示文稿提供了辅助描述。然而，这些解释往往被粗心的读者忽略，在呈现时观众看不到。Liu 等人[22]在写作阶段有效地生成有意义的文本，并写入到 PPT 文档的注释页中，以获得隐写文档。其中，生成的文本不仅与 PPT 文档的正文文本密切相关，而且还模拟了正文文本的写作风格。

针对 PDF 文档，Zhong 等人[23]通过分析 PDF 文档的结构特征可知，交叉引用表是 PDF 文件中唯一具有固定格式的部分。利用交叉引用表的行末标识符不会在文档中显示这一特性，可将其作为嵌入点。行末标识符有两种组合方式："\n"或者"\r\n"，根据秘密信息逐个修改行末标识符，例如"\n"代表秘密数据"0"，"\r\n"代表秘密数据"1"，实现秘密信息的嵌入。该方法能够有效抵抗统计检测的攻击，同时保护 PDF 文档。

（2）基于文本内容的方法

由于具有相同含义的内容可以用不同的方式表示，所以在嵌入秘密信息时，正确地改变表达式并不会损害其内部含义，能够实现在保持原文意思不变的基础上，将秘密信息嵌入到载体文本中。根据对载体文本修改的粒度不同，基于文本内容的方法旨在通过对文本中词汇级别和句子级别的语义单元进行替换后，仍可以保持文本的意思不变。

词汇级别的方法，最为典型的是同义词替换。例如，可以将"购买"替换为"买入"或者"采购"，将"快乐"替换为"愉悦"或"开心"，这些替换不会对原文本的语义造成很大的影响，但可以有效地嵌入秘密信息。同样地，在传输秘密信息时，通信过程中可以使用同义词代替敏感词汇，以此来避免敌人的窃听与破解。最早的同义词替换隐写系统[24]利用两对同义词嵌入秘密信息。Bolshakov[25]通过建立一个同义词库，其中包含绝对和相对的同义词，实现了同义词之间替换

效果的文本隐写方法。甘灿等人[26]采用依存句法分析，更加全面地获得同义词和搭配词之间的关联性。不同于 Bolshakov 只考虑相邻单词之间的相关性，该方法考虑到了在文本中距离较远单词也可能存在影响。Chang 等人[27]使用 N-Gram 模型来检查同义词在上下文中的合理性，同时开发了一种新的顶点编码方法。该算法利用图这一数据结构提高了文本的隐蔽性。

此外，还有一些方法从句子级别入手，利用句中词语之间的依赖关系，试图改变句子的结构，以消除部分词语的替换造成的突发性。在自然语言中，通过对句式结构进行调整能够做到句义不变的效果。Topkara 等人[28]利用主被动变换隐藏信息，接收方可以通过识别接收到隐写句子的主动或被动形式提取出对应的秘密信息。接收到的文本为主动句式可以表示秘密位“0”，而被动句式可以表示“1”。Murphy 等人[29]利用句法分析，对从句结构进行变换，这些转换不仅可以隐藏信息，而且不改变文档的含义或风格。Chang 等人[30]提出了一种基于词排序技术的文本隐写方法，通过改变单词顺序生成流畅的文本，然后建立生成的文本和所嵌秘密信息间的映射关系。

基于文本内容的方法可以高效地保持原始文本的语义内容而不进行改变，因此具有较高的稳健性。尽管基于内容的文本隐写方法在隐蔽性方面表现优异，但文本中同义词的替换数量是有限的，所以该方法的嵌入率较低。与基于格式的方法相比，基于内容的方法需要利用更深层次的文本线索，例如语义、语法和上下文等，来达到信息隐藏的目的，这也限制了该方法增加嵌入容量的能力，同时使得其嵌入过程变得更为复杂。

1.3.2 生成式文本隐写

生成式文本隐写是根据秘密信息和经过训练的语言模型直接生成载密文本的过程。在载密文本生成过程中，经过训练的语言模型将为单词池中的每个候选词分配一个预测概率，同时根据秘密信息进行匹配，这样当前输出的单词是“最合适”的，能够更好地保持文本质量。基于生成的文本隐写方法摆脱了替换规则的限制，因此可以将秘密信息嵌入到生成的每个单词中，从而与修改式文本隐写方法相比可以提供更高的有效载荷。

相较于修改式文本隐写方法，通过引入深度学习技术，生成式文本隐写方法在文本质量和可读性上均得到了一定的提升。因为生成式文本隐写方法能够做到在每个单词位置上嵌入秘密信息，因此嵌入容量较大。但该方法在生成过程中是依据秘密信息来选择生成词，因此可能会出现选择到的单词预测概率很小的情况，从而影响生成文本的流畅性。

早期生成式文本隐写算法主要利用马尔可夫模型产生载密文本。Dai 等人[31]提出了一种基于马尔可夫模型和美国数据加密标准（Data Encryption Standard, DES）算法的拼写语言文本隐写系统。该系统首先对原始消息进行编码并转换为文本序列，然后使用 DES 算法加密该序列并隐藏在载体文本中，最终，通过解密过程成功提取嵌入的消息。针对该算法没有考虑状态转移图中不同单词间存在概率大小的问题，Moraldo 等人[32]改进编码方式，将转移图中的状态根据概率进行编码，以此实现每个码字与短句之间的唯一映射关系。虽然这种方法考虑到了状态转移概率这一要素，从而提高了文本质量，但会损害嵌入有效载荷。因此，一些研究者开始关注特殊文本体裁，如中文中的诗歌、诗词等，以提高生成文本的质量。例如，Luo 等人[33]提出了一种新的基于马尔可夫模型生成中国古典诗体词的文本隐写术，该方法利用声调特征选择同属性的单词，解决了语义和语法的缺陷问题，并适合于中文诗的生成。但是，这些使用基于马尔可夫模型的方法中都存在着相同的一个困境。由于只考虑上下文相关的少数单词，所生成的文本的可读性较差，容易被基于统计的隐写分析工具检测出来[34]。因此，这些方法仍需要面对可读性与安全性之间的平衡问题。

近年来，随着自然语言处理技术和深度学习的快速发展，神经网络模型已渐渐取代统计模型被广泛应用在文本隐写中。在每个生成步骤中，发送方根据秘密信息、概率分布和编码算法选择一个唯一的单词。具有大量参数的深度学习模型在经过大规模高质量语料库训练后，在生成文本的过程中，可以保留更具有代表性的特征信息。引入深度学习的方法，不仅大大降低了生成看似自然文本的难度，也降低了引起攻击者怀疑的可能性。

有一些研究人员将循环神经网络（Recurrent Neural Network, RNN）[35]及其变体引入到文本隐写中作为生成语言模型。2017 年，Fang 等人[36]最先提出基于长

短期记忆网络(Long Short Term Memory, LSTM)[37]的文本隐写方法。该算法利用共享词典和块编码的方式对每个词进行编码，并通过秘密比特与 LSTM 的概率转移关系选择合适的词汇从而生成载密文本。相较于传统的马尔可夫模型，该算法显著提高了生成文本的隐蔽性。Yang 等人[38]进一步改进，基于 LSTM 网络基础上采用了定长编码方法和变长编码方法。由于动态映射关系，接收方需要获得相同的语言模型并重现生成过程。接收方需要计算每个时间步长的条件概率，致使该方法将消耗更多的时间来提取秘密信息。Kang 等人[39]引入主题词和注意力机制与 LSTM 网络相结合的方法，有助于提升句子间的连贯性，确保生成的文本与主题相符。虽然该方法有效地增加了使用场景，但在高有效载荷下，生成的载密文本存在一定程度的失真问题。

随着语言模型的发展，Transformer[40]结构进一步提高了生成文本的质量。当前自然语言处理领域两大先进的模型都采用了基于 Transformer 架构，BERT(Bidirectional Encoder Representations from Transformers)[41]的网络结构类似于 Transformer 的编码器(Encoder)部分，而 GPT(Generative Pre-trained Transformer)[42]类似于 Transformer 的解码器(Decoder)部分。Ziegler 等人[43]提出了一种基于算术编码的文本隐写算法，该算法注重其抗统计分析的性能。研究团队使用 GPT-2[44]作为文本生成模型，根据获得的概率分布构建同心圆，利用算数编码找到唯一的一条路径，该路径上的词语构成了载密文本，其中路径上各单词之间具有紧密联系。值得注意的是，该算法生成的载密文本在统计分布上与自然文本高度相似。语言模型的改进使得文本质量得以提高。Yi 等人[45]提出了一种基于 BERT 和吉布斯采样的位置驱动生成隐写算法，该算法将秘密信息嵌入到一个单词中，而不是二进制字符串。在这种方法中，嵌入位置需要在双方之间共享，实验结果表明，该算法能够达到目前较好检测性能。

生成式文本隐写方法通过引入深度学习技术在语义连贯性和文本质量方面得到了提高。同时，该方法在每个单词位置都可以嵌入秘密信息，因此可以提供较高的嵌入有效载荷[46]。但与修改式文本隐写方法相比，其安全性不能很好地被保证[36, 47-49]，以及在生成过程中依据秘密信息来选择生成词，可能会出现选择到的单词预测概率很小的情况，从而影响生成文本的流畅性。

1.4 研究内容与结构安排

1.4.1 主要研究内容

本文主要研究语言模型驱动的文本隐写。文本是日常生活中信息交换的重要媒介之一，文本隐写通过在文本中嵌入秘密信息来达到隐蔽通信的目的。本章对文本隐写的发展现状进行了详细的梳理，其中，如何生成携带秘密信息的高质量文本是一个关键问题。现阶段主流的文本隐写可以分为两类：修改式文本隐写和生成式文本隐写。修改式文本隐写主要针对文本格式或文本内容进行修改，虽然该方法隐蔽性好，但其嵌入有效载荷较低。生成式文本隐写是在文本生成过程中通过对候选词的选择实现秘密信息的嵌入。这类方法可以提供更高的嵌入有效载荷，但其安全性不能很好地保证。本文具体内容如下：

(1) 基于 BERT 和一致性编码的自回归文本隐写

本文提出了一种基于 BERT 和一致性编码的自回归文本隐写算法，旨在解决主流的非自回归文本隐写方法的不足。该算法在保证安全性的同时，提高了文本的嵌入有效载荷。通过掩码语言模型和一致性编码，我们需要对候选词集的尺寸设限，并利用单词的概率分布进行信息隐藏。这种方法充分考虑了上下文的联系，从而保持文本的质量。与之前的工作相比，实验结果表明，该算法在嵌入容量和系统安全性之间取得了更好的平衡，并且提高了载密文本的可读性。

(2) 基于掩码语言模型的可逆文本信息隐藏

针对现有文本信息隐藏方法中存在载体文本无法被完全恢复和双方共享的边信息较多等问题，本文提出了一种基于掩码语言模型的可逆文本信息隐藏算法。该算法基于掩码语言模型进行操作，以生成满足要求的载密文本，使得载密文本中的秘密信息可以被成功提取，并且通过收集某些位置的单词完美地重建载体文本。实验结果表明，该算法可以在保证文本流畅性和语义质量的同时，可以达到较好的安全性。此外，该算法不要求数据隐藏者和数据接收者共享同一个语言模型，这减少了彼此之间共享的边信息，具有更广的应用范围。因此，本算法通过使用掩码语言模型来实现可逆文本信息隐藏，优化了嵌入和提取的效果，兼顾了不可感知性、嵌入容量和可靠性。

1.4.2 论文结构安排

本文各章内容安排如下：

第一章，阐述了课题的研究背景与意义，介绍了文本隐写国内外研究现状，对修改式文本隐写和生成式文本隐写研究现状进行具体分析，并概括了本文主要的研究成果和论文组织结构。

第二章，介绍了文本隐写技术基本框架、方法概述和评价指标，阐述了文本隐写分析的基本概念、方法概述和评价指标。此外，根据语言模型的发展情况这一线索进行展开，并对 BERT 语言模型和内部结构进行了着重介绍。

第三章，提出了一种基于 BERT 和一致性编码的自回归文本隐写方法。首先阐述相关工作和该方法的整体框架，并对掩码策略、掩码语言模型和编码方式分别进行阐述，最后通过与其他算法进行定性与定量实验分析，并进行消融实验以有效验证基于 BERT 和一致性编码的自回归文本隐写方法的有效性与优越性。

第四章，提出了一个基于掩码语言模型的可逆文本信息隐藏方法。首先阐述该方法的整体框架，并对语义初始化、语义控制、数据嵌入、数据提取和载体重建五个部分进行阐述，最后通过与四种有代表性的信息编码策略进行实验分析，以有效验证基于掩码语言模型的可逆文本信息隐藏方法的有效性与优越性。

第五章，对本文的研究内容进行了总结，并对下一步研究进行了展望。

第二章　文本隐写相关技术介绍

2.1 文本隐写技术

2.1.1 文本隐写基本框架

隐写是一门古老的技术，古希腊时期是隐写技术发展的较早阶段。例如，人们在奴隶头骨上刻上秘密信息的纹身，等待他们的头发长出覆盖住纹身后再让奴隶旅行传递消息。在古希腊时期，人们还常常运用特殊墨水、蜡板和首字母符号等手段来隐藏秘密信息[50]。在现代社会，文本作为信息交流和传递的一种基础形式，广泛应用于人们的网络生活中。无论是日常聊天、电子邮件还是各种网站上的留言评论等，几乎都离不开文本这个工具。在这种背景下，如何利用文本为载体，安全地传输秘密信息成为十分重要的问题。

文本隐写，就是将秘密信息嵌入到载体文本中，实现对信息的隐藏和传输。现代的文本隐写依赖于文本的冗余性原理进行嵌入，文本隐写的基本框架是基于 Simmons 提出的“囚犯问题”的场景[51]进行描述，如图 2.1 所示。Alice 和 Bob 被关在了相距很远的两个单独牢房里面，他们为了一起合谋完成越狱行动，彼此之间需要相互通信，传递秘密信息。但是看守者 Wendy 会监督并检查他们的通信文件，并决定是否能够完成信件的传递。

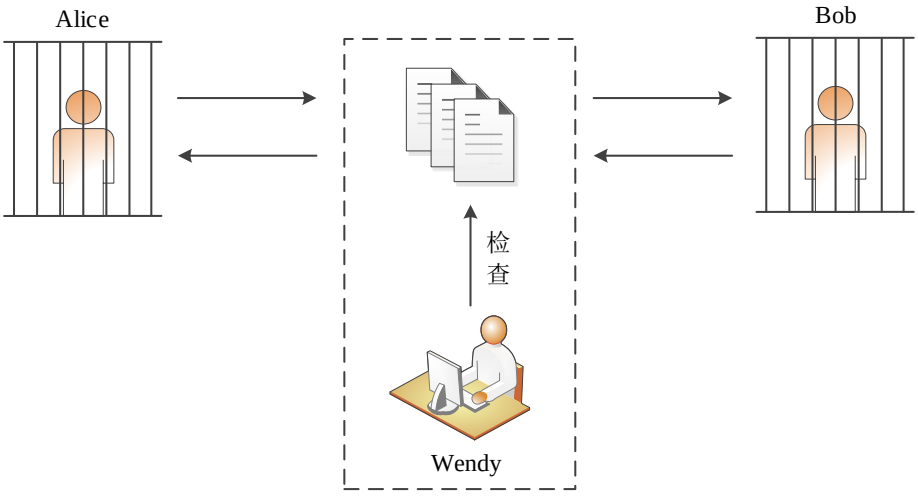


图2.1　囚犯问题[51]

为了 Alice 和 Bob 需要在不被看守者 Wendy 发现的情况下秘密地进行交流，研究人员将“囚犯模型”这一理想化的模型一般化，其关键问题在于如何在原始载体中嵌入秘密信息，即 Alice 的数据嵌入，以及如何从包含秘密信息的隐写载体中提取秘密信息，即 Bob 的数据提取。为此，研究人员设计了如图 2.2 所示的文本隐写系统来完成秘密通信的过程。一般来说，文本隐写的基本框架可以分为信息嵌入和信息提取两部分。

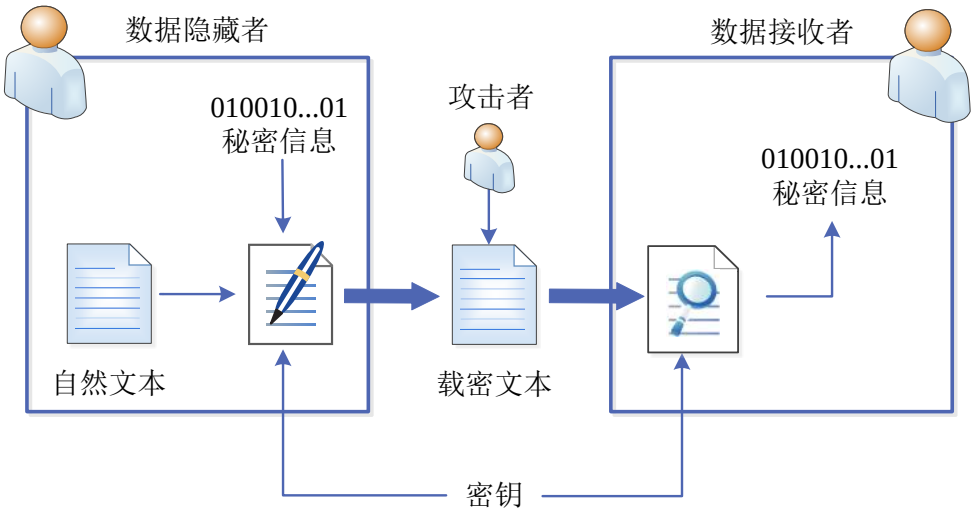


图 2.2　文本隐写的基本模型

如图 2.2 所示，为了将秘密信息隐藏起来，数据隐藏者可以运用嵌入算法和密钥，将秘密信息和载体文本进行结合，形成一个载密文本，这就像“囚犯问题”中 Alice 所面临的任务一样。通常情况下，秘密信息会预先转换成二进制的形式为了更方便地嵌入。在不安全信道的传输过程中，面临着被第三方攻击者检测的困境，与“囚犯问题”中 Wendy 所遇到的挑战相似。数据接收者需要使用提取算法和共享的密钥来解密载密文本，提取原始的秘密信息，这就像“囚犯问题”中 Bob 所面临的任务一样。

然而，文本作为一个载体有其固有的局限性，可供存储秘密信息的冗余空间较小，因此需要更为高效的文本隐写算法来保证载密文本的不可感知性。文本隐写技术的难度主要体现在两个方面：一是如何尽量隐蔽地嵌入秘密数据，即避免文本出现明显异常的情况；二是如何防止相关攻击，即对文本进行修改或提取秘密信息。因此，研究文本隐写技术的过程可以促进信息安全领域的发展，使信息处理更加安全和高效。

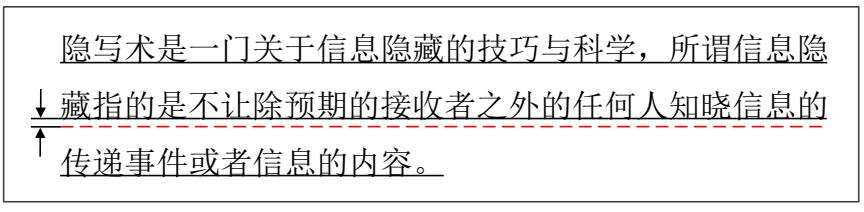
当前，文本隐写主要分为两类：修改式文本隐写和生成式文本隐写。前者通常通过修改载体文本的格式和内容的方式嵌入秘密信息，而后者则根据秘密信息和经过训练的语言模型直接生成载密文本。利用语言模型和信息编码方式直接生成包含秘密信息的文本。各种文本隐写技术形式多样、复杂，下面将对文本隐写的典型方法进行展开介绍。

2.1.2 文本隐写方法概述

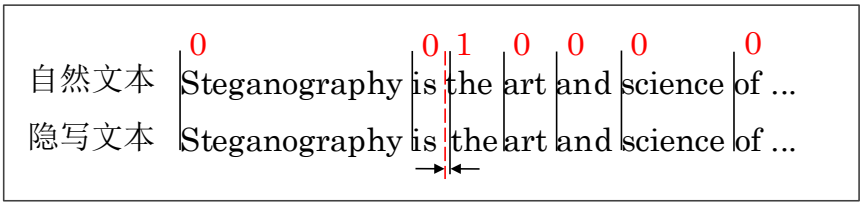
（1）基于文本格式修改

基于文本格式的隐写主要利用文本或文档的结构和格式特点，通过修改多个文本中的格式，如间距、字体大小和字体颜色等，来隐藏秘密信息。在该方法中，一般使用二进制数来表示秘密信息，以便更有效地实现信息的嵌入。下面将介绍基于文本格式修改的文本隐写方法，并且提供示例以帮助您更好地理解该技术。图 2.3 是两种基于文本格式的文本隐写算法。

针对文本的间距这一格式特性，Low 等人[52]提出了行移编码和字移编码算法。如图 2.3(a)所示，行移编码是一种通过向上或向下移动文本中的几行来嵌入秘密信息，如果改变了行间距，则嵌入秘密位“1”，否则，如果行间距不改变，则嵌入秘密位“0”。同样，如图 2.3(b)所示，字移编码是通过向左或向右移动字符在文本行中的位置来修改文本，改变字符间距与行间距类似。该方法应用场景为相邻字符间间距可变的文档时，该编码算法隐蔽性较好。



(a)



(b)

图2.3 基于文本格式的隐写算法示例

而修改字体大小、字体颜色等其他文本格式的方法与上述方法之间存在相似性，只是修改的文本格式不同。虽然这种方法实现简单，但也意味着更容易被格式检查类的隐写分析工具发现，从而造成信息泄漏。

（2）基于文本内容修改

相较于基于格式修改的文本隐写方法，基于文本内容的隐写方法有所不同，它会在保持文本格式不变的前提下，对文本内容进行修改以隐藏秘密信息。最为典型的方法是同义词替换[26, 53-55]，其过程通常经历以下几个步骤。首先构建一个包含同义词表的字典，例如WordNet[56]或者中文的《同义词词林》[57]，对同义词词典中的词汇进行适当的编码，接着，发送方选择与待嵌入秘密信息相对应的同义词进行替换，从而生成载密文本。需要注意的是，为了保证信息的安全性，同义词词典和编码方式可能需要采用一些随机化或加密手段。

如图 2.4 展示了基于同义词替换的文本隐写方法。对于自然文本“*Midshire is a wonderful little city.*”，假设数据隐藏者想要传递秘密信息“101”，经过同义词替换后，得到的载密文本为“*Midshire is a nice little town.*”。虽然秘密信息可以通过同义词替换成功嵌入，但部分词替换后不能保证整个句子的流畅性。因此，在实际使用过程中，需要精细的同义词处理策略来保证信息的准确性和合理性。

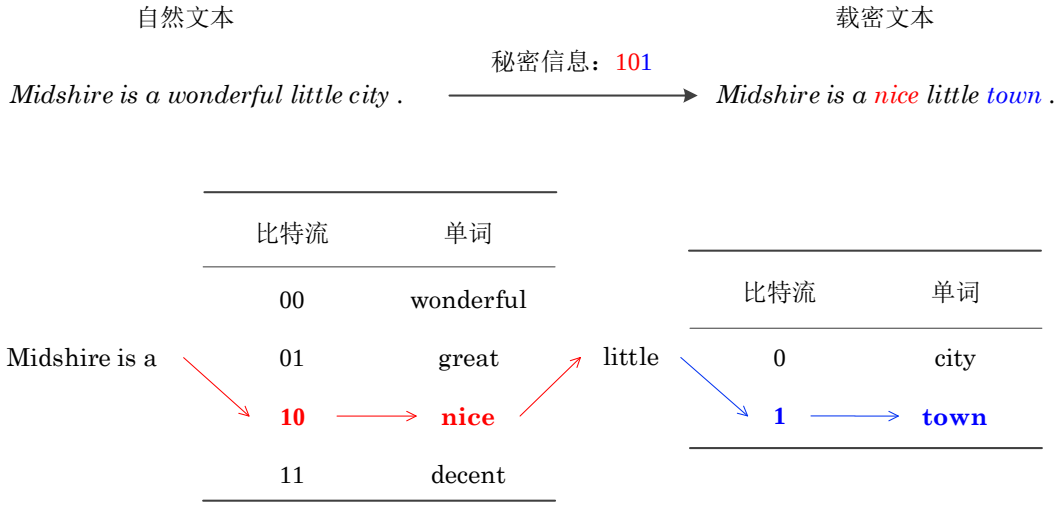


图2.4　基于同义词替换的文本隐写方法

（3）基于生成的文本隐写

早期的基于生成的文本隐写方法主要包括基于固定模板的方法，这种方法在一定程度上能够保证文本的语法正确性，但会导致文本上下文的语义偏离，从而

不利于秘密信息的隐藏。随着深度学习技术的发展，研究人员开始使用循环神经网络、Transformer 模型及其变体等语言模型来生成载密文本，采用这些先进模型可以更好地提高文本的质量和实现更佳的信息隐藏效果。通常情况下，基于生成的文本隐写方法具有以下三个主要步骤。第一步：选取数据集，并利用语言模型进行训练，同时在迭代过程中计算候选词语出现的概率；第二步：根据算法设计规则建立候选词语的集合，利用特定的信息编码技术实现候选词语与二进制序列之间的映射关系；第三步：通过映射关系选择相应的候选词语，直到所有的秘密信息比特流都被完全嵌入至载体文本中。

基于生成的文本隐写不仅能够完成高质量的秘密通信任务，而且在语义上也更加准确，提高了秘密信息的隐蔽性。该方法的核心部分是数据嵌入，通常涉及到文本生成和信息编码这两个重要的步骤。前者主要依赖于自然语言处理中应用的各种文本生成技术和语言模型，本文在 2.3 节中对语言模型的发展进行了详细描述。后者涉及各种设计良好的信息编码技术，其目标是建立单词和秘密位之间的映射关系。进行数据嵌入后，将得到的载密文本发送给数据接收方，数据接收方根据密钥和必要的边信息提取隐藏信息。这些编码方式会对载密文本的质量产生影响，因此，下面将对常见的信息编码策略进行介绍。

（a）块编码

给定一组候选词，块编码方法根据预测概率从该组中选出的 2^k 个单词分配到从 0 到 2^k-1 的 k 位二进制码中[58]。例如，载体文本为“*We finish the charitable project.*”。同义词集{*complete*，*finish*}两个单词，则 $k=1$，因此 1 位二进制代码“0”和“1”分别被分配给 *complete* 和 *finish*。因为同义词集{*labor*，*project*，*task*，*undertaking*}四个单词，则 $k=2$，所以块编码方法可以使用 1 位或 2 位二进制数对单词进行编码，如图 2.5 所示。当使用一位块编码时，*labor* 和 *task* 都表示“0”，*project* 和 *undertaking* 都表示“1”；当使用两位块编码时，这四个单词被分配不同的二进制数“00”、“01”、“10”和“11”。

假设秘密是一个随机二进制字符串，使用 1 位进行编码的优点是，载体文本中的单词 *project* 只有 50%的概率需要替换为它的同义词，而 2 位进行编码的方案有 75%的机会修改载体单词。然而，1 位块编码方式嵌入的信息更少。因此，

在安全性和有效载荷能力之间存在一种权衡。值得注意的是，在这个简单的方案中，每个块编码表示都有相同的被选择的概率。但是对于母语使用者而言，同义词选择中会存在倾向性，而不是存在同等被选择的可能。

	编码	单词
We	0	*complete*
	1	*finish*

	编码	单词
the charitable	00	*labor*
	01	*project*
	10	*task*
	11	*undertaking*

图2.5　块编码的示例

(b) 霍夫曼编码

霍夫曼编码[59]是可变字长编码的一种编码方式，在基于生成的隐写算法中被多次使用到，这种编码方式根据单词出现的概率来构造平均长度最短的码字，这样可以有效地减少冗余编码。图 2.6 展现了如何在隐写系统中使用霍夫曼编码完成的编码过程，预先假设每个候选词语都有一个出现的概率分数，霍夫曼算法确定了一种为每个单词分配一个可变长度的二进制字符串的方法，可以看到，出现概率越高的单词转化后的二进制编码就越短，也就是说，母语使用者经常使用的词语更有可能被隐写系统所选择。

	编码	单词	概率
We	0	*complete*	0.77
	1	*finish*	0.23

	编码	单词	概率
the charitable	110	*labor*	0.05
	0	*project*	0.69
	10	*task*	0.25
	111	*undertaking*	0.01

图2.6　霍夫曼编码的示例

图 2.7 是根据候选词的概率分数构建霍夫曼树这一过程，每个叶子节点都包含一个单词及其相关的概率。从根节点开始，将其左孩子节点编码成“0”，右孩子节点编码成“1”，利用深度优先搜索算法得到秘密信息对应的叶子节点单词。其中具有较低概率的候选词在树的底部，较高概率的候选词在树的较高处，也就是说，概率分数越高的单词映射的二进制编码就越短，这确保了更“合适”的单词更有可能被选择。

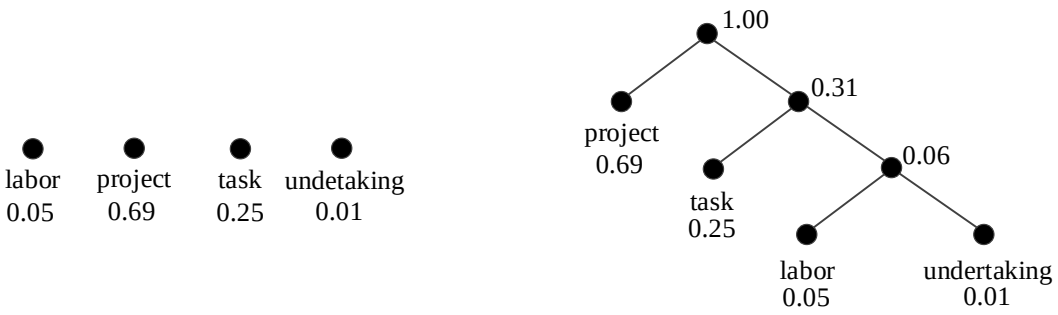


图2.7　构建霍夫曼树的过程[59]

因此，对于隐写系统而言，双方实现共享信息编码方式后，接收方仍然能够唯一地解码发送方编码后得到的载密文本。例如，载密文本为“*We finish the charitable project.*”，则可以唯一解码得到秘密信息为“10”。在每一个步骤中，选取单词的概率大于阈值的候选词，利用霍夫曼进行编码，这样在高概率的单词中进行选择，这最终输出的词语更符合母语使用者的习惯。

2.1.3　文本隐写评价指标

随着自然语言处理技术的不断发展，与之相对应的检测技术应运而生，也就是，用来分析传输的载体文本中是否嵌入秘密信息的方法。因此，了解文本隐写的评价指标，对文本隐写方法的突破能产生巨大的作用和影响。一个文本隐写系统必须满足两个基本要求[27]。

（1）嵌入容量

第一个要求是衡量载体文本中嵌入容量的能力。嵌入容量，即有效载荷，是指数据隐藏者在文本的冗余空间中所能传输的秘密消息的容量。在保证隐蔽性的条件下，通常希望在载体文本中嵌入尽可能多的信息，即一次传递更多的秘密。为了计算嵌入信息的数量，比特每个词(Bits Per Word, BPW)已被广泛应用于文本隐写方法中。有效载荷的计算方法如公式(2.1)所示：

$$\mathrm{BPW}=\frac{L_c}{L_r}\times 100\% \tag{2.1}$$

其中，L_c 表示秘密信息的比特数，L_r 表示载体文本的长度，两者做商所得即为嵌入有效载荷数。在秘密信息安全传输的情况下，嵌入有效载荷越高，则该隐写算法所能传输的秘密信息越多，效率越高。

（2）不可感知性

第二个重要的要求是不可感知性。文本隐写系统最重要的目的是隐藏通信的存在，以确保载密文本的不可怀疑性。这意味着当载密文本被人类或计算机分析时，它不应该揭示隐藏信息的任何证据。换句话说，人类的感知系统和隐写分析器都不可能区分载密文本与自然文本。因此，不可感知性可以从两个方面来衡量的，即感官不可感知性和统计不可感知性。

感官不可感知性指的是人们无法通过感觉来区分载密载体和自然载体。为了我们需要确保文本在被嵌入秘密信息后仍然是流畅的、容易理解的，一般存在两种评价方法，一是主观评价方法，主要根据评价人员在感官上对载密文本和自然文本进行打分来衡量，不论是视觉、触觉还是听觉都无法感知到信息的存在，不会引起攻击者的怀疑；二是客观评价方法，根据机器来客观评价载密文本的质量，通常使用困惑度（Perplexity, PPL）[60]这一指标来定量分析文本的可读性。设载密文本为$S=\{w_1,w_2,\ldots,w_n\}$，模型所生成单词w_i的概率为$P(w_i)$，则 PPL 计算方法如公式（2.2）所示：

$$\mathrm{PPL}=2^{-\frac{1}{n}\log P(w_1,w_2\ldots,w_n)} \tag{2.2}$$

统计不可感知性是所获得的隐写载体需要有能够抵抗隐写分析的能力，意味着确保隐写载体和自然载体在统计分布上不可区分、相似度高。为了衡量这种统计不可感知性，需要进行与隐写相对应的隐写分析实验，这部分将在 2.2 节中进行详细介绍。文本隐写分析能够辨别传输的载体文本中是否隐藏了秘密消息，并采取相应措施进行鉴别。

以上，介绍了文本隐写技术的相关评价指标，主流的文本隐写方法侧重于提高其嵌入容量和不可感知性。由于隐写术的目的是秘密信息的传输，因此它需要足够的嵌入能力。理想的隐写系统应具有较高的不可感知性和较大的有效载荷能力。然而，它们之间的关系是相互竞争和矛盾的[27, 61, 62]，因为任何试图在载体媒体中嵌入额外信息的操作都可能增加向数字媒体中引入异常的机会，从而降低不可感知性。不可感知性的增强将减少嵌入的有效载荷，而嵌入的有效载荷的增加将需要对原始文本进行更多地修改，从而降低不可感知性，也导致安全性降低。因此，如何在它们之间实现良好的平衡是文本隐写的一个核心问题。

2.2 文本隐写分析技术

2.2.1 文本隐写分析基本概念

随着文本隐写术的不断提升，与之对应的隐写分析技术也不断发展，两者是相互对抗的。文本隐写分析能够辨别传输的载体文本中是否潜藏了秘密消息，并采取相应措施进行鉴别。如果某种文本隐写算法具有卓越的抗隐写分析能力，则可以认为该算法的安全性较高。反之，如果算法的抗隐写分析能力较弱，则该算法很容易被侦测或攻击，安全性则较低。

隐写分析是一种用于检测、识别和解读隐写信息的技术[63]。为了区分自然文本和载密文本，文本隐写分析算法通过统计载体文本的特征变化来辨别。这些特征变化可以是在词汇、语法或其他方面的差异，从而帮助算法判断是否存在隐藏信息。主流的文本隐写分析可以建模成一个二分类任务，假设隐写分析器检测到载体文本中含有秘密信息，则隐写分析结果为真，表示存在秘密信息；相反，若隐写分析结果为假，则该载体为自然文本，即没有嵌入秘密信息。

2.2.2 文本隐写分析方法概述

文本隐写分析方法主要分为人工特征提取和自动特征提取两类。早期的研究集中在手工特征提取，这通常被称为传统的文本隐写分析，而目前的研究集中在自动特征提取，这可以被称为端到端文本隐写分析。

（1）传统的文本隐写分析

传统隐写分析的核心思想是使用统计机器学习算法对信息嵌入引起的细微差异进行建模和检测，从而识别载密文本。传统的文本隐写分析方法主要分为目标隐写分析和盲隐写分析两类。

目标隐写分析是一种用于针对特殊文本隐写算法的隐写分析方法。该检测算法需要事先知道使用哪种文本隐写方法来嵌入秘密信息，然后分别设计相应的特征。因此，目标文本隐写分析更擅长于检测特定的文本隐写算法。但是，当面对其他隐写算法时，该隐写分析的准确率可能会出现指数级别的下降。隐写分析算法通常使用统计分析检测秘密信息的存在，在目标文本隐写分析方法中，比较常

见的统计特征主要包括邻域差分(Neighbor Difference)[64]、单词首字母分布(Word-initial Distribution)[65]、字体属性(Font Attribute)[66]和同义词频率(Synonym Frequency)[12, 67]。

除了上述针对特定隐写算法的目标隐写分析方法，研究者们也提出了许多通用的隐写分析算法，也就是盲隐写分析。由于在普通文本中嵌入秘密信息或多或少地改变了文本的内容，从而引入了普通文本特征的统计差异，如单词的流畅性和分布。例如，Meng 等人[68]提出了一种基于统计语言模型的隐写分析算法，该算法使用 PPL 值来确定给定的文本是否包含秘密信息。实验结果表明，在足够的训练数据下，检测精度良好。Chen 等人[69]提出了一种基于单词分布的盲文本隐写分析统计算法。该算法以单词分布作为特征向量进行分类，并以 SVM 作为分类器。Zhao 等人[70]利用单词的类信息熵统计变量及其方差作为两个特征来提高检测精度。因此，这类方法满足了更广泛的应用和要求。

(2) 端到端的文本隐写分析

传统的语言隐写分析方法要求研究人员手工设计各种启发式特征，以揭示文本隐写的痕迹。一方面，这种文本分析需要研究者仔细设计相应的特征，检测精度主要取决于研究者的经验知识。另一方面，随着文本隐写的发展，先进的隐写方法可能不会在载密文本中引入明显的统计差异，因此这些隐写方法很难被发现。因此，迫切需要探索新的隐写分析方法。研究人员引入端到端的文本隐写分析，来获取更有效的特征发现隐写的证据。

在深度学习刚被引入文本隐写分析的时候，研究人员受到传统的文本隐写分析方法的影响，认为仅使用简单的深度学习网络结构去提取文本本身比较明显的特征就可以实现较好的检测效果。Yang 等人[71]提出了一种基于语义分析的文本隐写分析方法，最先使用卷积神经网络(Convolution Neural Network, CNN)提取文本的高级语义特征，并将其映射到高维语义空间。Li 等人[72]实现了一种基于胶囊网络的文本隐写分析方法，用胶囊代替神经元。实验结果表明，该方法能够提取和保留更丰富的文本语义特征，从而实现有效的隐写分析。

自然文本具有复杂的句法结构、丰富的依赖性和单词共现信息。众所周知，单词之间的相关性是全局的，不仅限于相邻的单词之间。因此，需要充分利用句

法结构、全局特征、上下文信息等各种特征来进行有效的隐写分析。为了解决这一问题，Niu 等人[73]提出了一种基于双向-LSTM 网络（Bi-LSTM）和非对称卷积核的文本隐写分析方法。他们使用 Bi-LSTM 来捕获文本的长期语义信息，并利用不同大小的非对称卷积核来提取单词之间的局部关系。Bao 等人[74]将注意机制融合到 LSTM-CNN 模型中，在获取局部和全局特征的同时，集中关注载体文本和载密文本之间最大差异的特征。Xu 等人[75]设计了一个特征交互模块来探索局部和全局语义特征之间的交互关系，该方法在面对不同语言风格的载密文本的场景中仍然表现良好。

端到端的文本隐写分析能够获取更有效的特征发现隐写的证据。一般来说，端到端的文本隐写分析不需要手工制作的特征，它通过基于深度学习的方法来构建特征。因此，在实际场景中应用端到端的文本隐写分析是更加实用的。

2.2.3 文本隐写分析评价指标

主流的文本隐写分析可以建模成一个二分类任务[76]，因此文本隐写分析的衡量指标与通常用于二分类任务的指标相同，包括准确率（Accuracy, Acc）、精准率（Precision, P）、召回率（Recall, R）和 F1 值（F1-score, F1）。文本隐写分析的主要目的是发现携带隐藏信息的文本，所以一般把携带隐藏信息的载密文本视为正样本。这些指标的计算方式如下：

$$\begin{aligned} Acc &= \left(N_{\mathrm{tp}} + N_{\mathrm{tn}}\right) / \left(N_{\mathrm{tp}} + N_{\mathrm{fp}} + N_{\mathrm{tn}} + N_{\mathrm{fn}}\right) \\ P &= N_{\mathrm{tp}} / \left(N_{\mathrm{tp}} + N_{\mathrm{fp}}\right) \\ R &= N_{\mathrm{tp}} / \left(N_{\mathrm{tp}} + N_{\mathrm{fn}}\right) \\ F1 &= \frac{2 \times P \times R}{P + R} \end{aligned} \tag{2.3}$$

其中，N_{tp} 表示正确分类的载密文本数，N_{fp} 表示正确分类的自然文本数，N_{tn} 表示错误分类的隐写载体数，N_{fn} 表示错误分类的自然文本数。如果计算出的 Acc、P、R 和 $F1$ 值越高，说明隐写分析的性能越好。这也意味着该方法能够更好地检测出载密文本，其所采用的隐写方法隐蔽性较差。文本隐写分析的评价指标能够对载密文本的不可感知性和隐蔽性进行定量评估。

2.3 自然语言模型

语言模型指的是一种用于自然语言处理的统计模型，它可以预测一个给定文本序列下一个词出现的可能性，即用来计算给定文本的总体概率分布的一种数学模型[77]。通过建立这样的数学模型，可以更好地理解和分析自然语言处理的相关工作。例如，对于一个由 n 个单词组成的句子 $s=\{w_1,w_2,\ldots,w_n\}$，只需要将其各个单词的条件概率相乘就可以计算出该句子的概率分布，具体计算方法如下：

$$\begin{aligned} P(s) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2 w_1)\ldots P(w_n \mid w_{n-1}\ldots w_1) \\ &= \prod_{i=1}^{m} P(w_i \mid w_{i-1}\ldots w_1) \end{aligned} \tag{2.4}$$

在当前主流的文本生成算法中，自回归式语言模型是被广泛采用的一种。自回归式语言模型是通过依次生成单词的方式，按照句子顺序来进行语句生成。其中，语言模型的选择对于预测下一个单词出现的概率以及生成文本的质量都有着至关重要的影响。语言模型主要分为两种：基于 N-Gram 语言模型的统计语言模型和基于神经网络的语言模型。本节将详细介绍几种常见的语言模型，为本文后续提供技术与理论的理解。

2.3.1 基于 N-Gram 的语言模型

N-Gram 模型将一段文本看作是若干个不重叠的 n 个单词序列，利用已有的训练数据来学习每个序列出现的频次以及其与其他序列之间的关系，从而计算任意一段文本的概率分布[78]。语言模型发展初期，统计语言模型快速发展，N-Gram 语言模型和马尔可夫模型的相似之处在于，它们都利用了前面出现的单词来计算当前单词的概率。具体而言，对于这类模型来说，第 i 个单词的预测概率是根据前面 $i-1$ 个单词决定的，在此基础上建立概率分布模型，即：

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) = P(w_i \mid w_{i-n+1} w_{i-n+2} \ldots w_{i-1}) \tag{2.5}$$

随着句子长度的增加导致计算量增大，为了缓解这个问题，N-Gram 语言模型引入了马尔可夫假设，则利用最大似然估计（MLE）的方法计算第 i 个单词的预测概率，如公式（2.6）所示：

$$P\left(w_i \mid w_{i-(n-1):i-1}\right)=\frac{\text{count}\left(w_{i-(n-1):i}\right)}{\text{count}\left(w_{i-(n-1):i-1}\right)} \tag{2.6}$$

其中，$\text{count}\left(w_{i:j}\right)$表示序列$w_{i:j}$在训练集语料库中出现的频次。

一元模型(Unigram)即当$n=1$时，表明任何一个单词的出现都只跟自身相关：

$$p\left(w_1,w_2,\ldots,w_n\right)=p\left(w_1\right)p\left(w_2\right)\ldots p\left(w_n\right) \tag{2.7}$$

二元模型(Bigram)即当$n=2$时，表明任何一个单词的出现都只跟前一个单词相关：

$$p\left(w_1,w_2,\ldots,w_n\right)=p\left(w_1\right)p\left(w_2 \mid w_1\right)\ldots p\left(w_n \mid w_{n-1}\right) \tag{2.8}$$

三元模型(Trigram)即当$n=3$时，表明任何一个单词的出现都只跟前两个单词相关：

$$P\left(w_1,w_2,\ldots,w_n\right)=P\left(w_1\right)P\left(w_2 \mid w_1\right)\cdots P\left(w_n \mid w_{n-1},w_{n-2}\right) \tag{2.9}$$

以上公式中的n可以根据具体的实验要求进行设置，n越大，意味着当前的单词与距离更远的单词产生依赖，联系更为密切，可以提供较多的上下文语境。但随之也伴随着模型参数增加，计算代价增大。

综上所述，尽管统计语言模型的发展已经相对迅速，但它仍然存在一些显著局限。首先，这类模型只能捕捉短距离依赖关系，影响了其生成文本的质量和可读性；其次，在解决自然语言实际应用的挑战时，概率计算误差往往不可避免，使得模型预测的准确率较低。

2.3.2 基于循环神经网络的语言模型

随着硬件技术逐渐普及以及数据集的不断优化，深度学习已经成为计算机领域中最热门的方向之一。在自然语言处理领域，研究者们开始使用深度神经网络(Deep Neural Network, DNN)[79]来作为语言模型提取特征信息。与以往采用人工方法进行特征提取的方式相比，深度学习通过监督学习的方式对数据进行特征抽取，主要利用层次化的神经网络结构完成。目前，深度学习已被广泛应用于众多领域，其中包括自然语言处理。

其中，循环神经网络(Recurrent Neural Network, RNN)[80]是自然语言处理领域最常用的语言模型，主要针对序列进行建模，其展开示意图如图 2.8 所示。正是 RNN 的时序关系的内部结构，使得它很适合被应用在序列数据的任务中，如文本、语音和视频中。

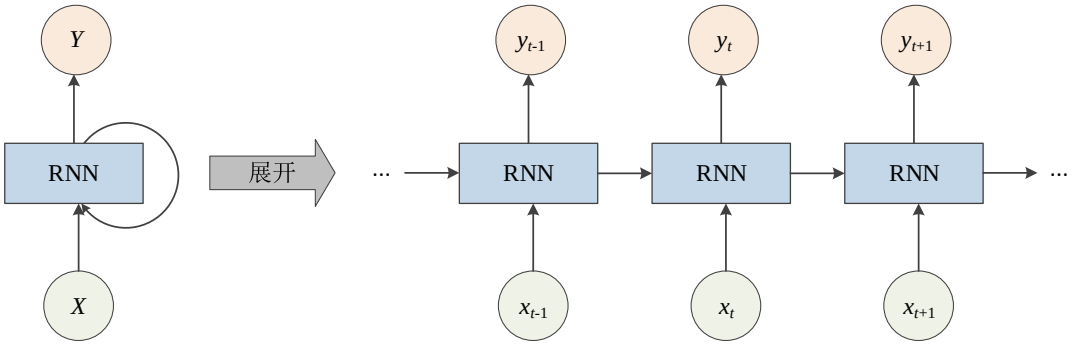


图2.8　RNN神经单元结构图

在 RNN 网络中，输入层的向量取决于前面所有词向量所包含的信息。相对于 RNN 的隐藏层，输出层的权重矩阵和激活函数是关键的两个不同之处。具体而言，输出层的激活函数使用了 Softmax 函数，但是隐藏层采用 Sigmoid。下列是相关计算公式：

$$x(t) = w(t) + s(t-1) \tag{2.10}$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right) \tag{2.11}$$

$$y_k(t) = g\left(\sum_i s_j(t) v_{kj}\right) \tag{2.12}$$

其中，$w(t)$表示前输入的词向量，$s(t-1)$表示前一个隐藏层向量。

虽然基于 RNN 的语言模型可以在高度连续的时间序列数据上表现良好，在处理长距离依赖关系方面表现出色，但是它仍存在一些限制。例如，很容易地面临梯度消失(Vanishing Gradient)或者梯度爆炸(Exploding Gradient)的问题。为了解决上述问题，研究者们在 RNN 基础上进行改进，其变体长短期记忆网络(Long Short Term Memory, LSTM)随之产生，在自然语言处理的各项任务上都取得了不错的成绩。与 RNN 相比，LSTM 在此基础上增加了遗忘门态(Forget Gate)和单元状态(Cell State)，LSTM 的单元结构示意图[81]如图 2.9 所示。

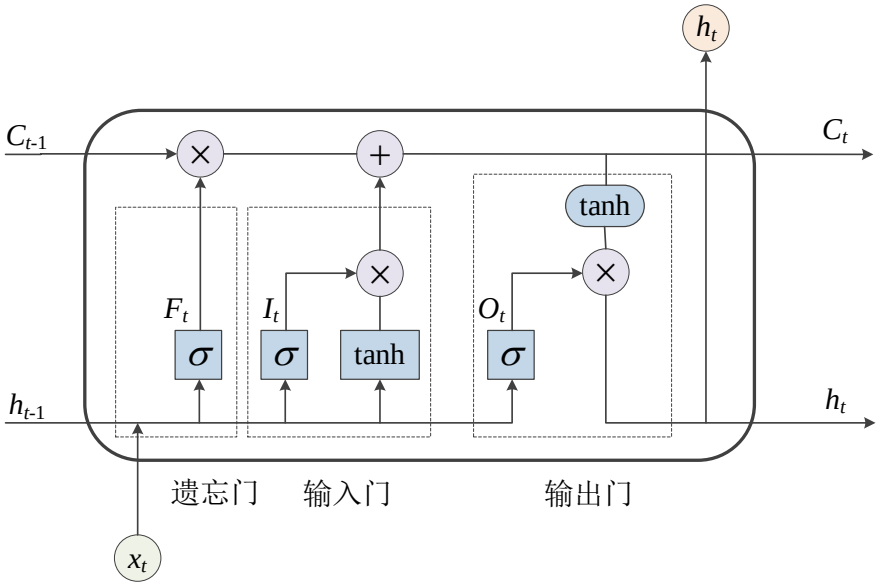


图2.9 LSTM单元结构示意图[81]

LSTM 更新状态如公式(2.13)所示，其中，$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示各个单元的权重矩阵和偏置值，均是需要训练得到的参数。$\boldsymbol{I}_t$，$\boldsymbol{F}_t$，$\boldsymbol{O}_t$ 分别表示输入门、遗忘门和输出门。$\boldsymbol{C}_t$ 表示记忆细胞，之前时刻的记忆信息保存在其中。σ 使用的是 sigmoid 函数，$\boldsymbol{h}_t$ 表示隐藏层的输出信息。因此，能够 LSTM 对序列建模的时候具有长期依赖记忆的能力。

$$\begin{cases} \boldsymbol{I}_t = \sigma\left(\boldsymbol{W}_i \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \\ \boldsymbol{F}_t = \sigma\left(\boldsymbol{W}_f \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right) \\ \boldsymbol{C}_t = \boldsymbol{F}_t \cdot \boldsymbol{C}_{t-1} + \boldsymbol{I}_t \cdot tanh\left(\boldsymbol{W}_c \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_c\right) \\ \boldsymbol{O}_t = \sigma\left(\boldsymbol{W}_o \cdot \left[\boldsymbol{h}_{t-1}, \boldsymbol{o}_t\right] + \boldsymbol{b}_o\right) \\ \boldsymbol{h}_t = \boldsymbol{O}_t \cdot tanh\left(\boldsymbol{C}_t\right) \end{cases} \tag{2.13}$$

在基于生成的文本隐写领域，最初的研究者采用基于 N-Gram 的语言模型来生成载密文本。然而，这种方法存在明显的局限性，因为它忽略了单词之间的长距离依赖关系。这就导致了生成的载密文本质量较差，不能够达到预期的效果。为了解决这个问题，后续的学者们开始引入 RNN 及其变体的语言模型，为此取得了显著进展。这些模型能够对序列数据进行处理，并将上一次的输出作为本次的输入，从而实现信息的传递和记忆。这些模型能够有效捕捉长距离的依赖关系，在生成载密文本方面表现优异，为该领域的研究带来新的动力。

2.3.3 Transformer 语言模型

不同于 RNN 模型的循环递归结构，Transformer 模型[40]采用自注意力机制和多头注意力机制相结合的结构。这种结构使得序列中每个位置都能直接与序列中所有位置相关联，从而更加准确、快速地捕捉序列之间的依赖关系。这种方法已经广泛应用于翻译、问答等自然语言任务，并取得了显著的成功。在训练阶段，RNN 模型需要在每一时刻计算当前隐状态的信息并传递给下一个时刻作为输入，而 Transformer 模型中的全局注意力机制则可以并行处理整个序列的信息，因此训练过程更快、更高效。

Transformer 模型是由编码器和解码器两部分组成，其中每个部分都堆叠了多个单独的 Transformer 基本单元。每个基本单元都包括四个部分：(1)多头注意力机制，用于捕获不同位置之间的关系；(2)残差连接，可以避免梯度消失问题；(3)层归一化，使数据更容易训练并提高准确性；(4)全连接网络，用于增强表示和决策能力。Transformer 模型具体结构如图 2.10 所示。

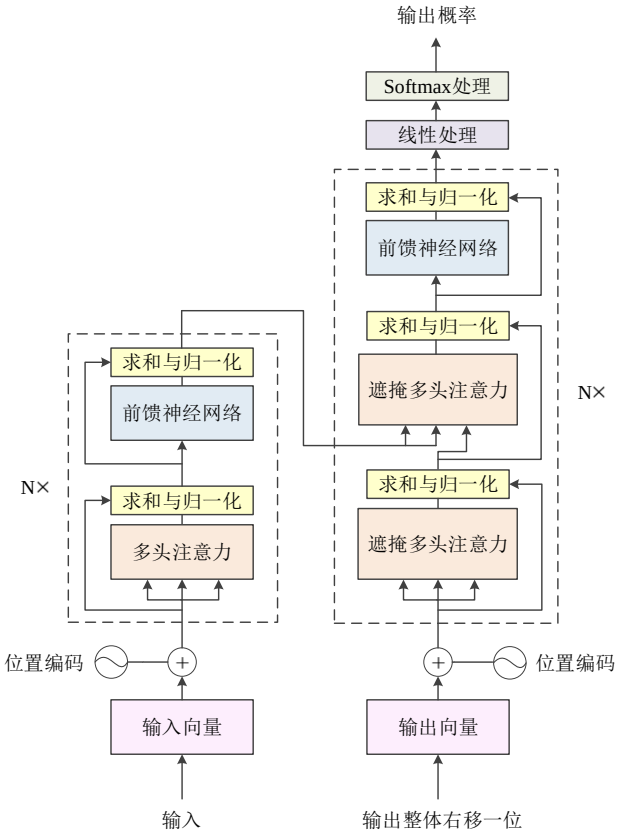


图 2.10 Transformer内部结构框架[40]

多头注意力机制(Mutli-Head Attention)是 Transformer 单元的核心，目的是更全面地抽取特征。多头注意力机制可以用以下公式进行描述：

$$MultiHead\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=Concat\left(\text{head}_1,\ldots,\text{head}_\text{h}\right)\boldsymbol{W}^{O} \tag{2.14}$$

其中，$\boldsymbol{Q}$，$\boldsymbol{K}$和$\boldsymbol{V}$分别代表查询、键和值的矩阵。h表示头的数量，$head_i$表示第i个头的输出，$\boldsymbol{W}^{O}$表示输出变换矩阵。每个头的输出$head_i$可以表示为：

$$head_i=Attention\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\ \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\ \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right) \tag{2.15}$$

其中，$\boldsymbol{W}_i^{\boldsymbol{Q}}\in R^{d_{model}\times d_k}$、$\boldsymbol{W}_i^{\boldsymbol{K}}\in R^{d_{model}\times d_k}$、$\boldsymbol{W}_i^{\boldsymbol{V}}\in R^{d_{model}\times d_v}$都是可学习的参数。在多头注意力机制中主要采用的是缩放点积注意力。缩放点积注意力实现了一个点乘注意力机制，通过$\boldsymbol{Q}$和$\boldsymbol{K}$点乘，再经过 Softmax 函数获得对$\boldsymbol{V}$的权重。最后进行加权和计算输出，具体如下公式所示，其中，d_k表示向量维度。

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2.16}$$

Transformer 是一种基于神经网络的语言模型，被广泛认为能够更好地捕捉文字信息，并且具有训练过程中更为稳定的特点，因此被誉为自然语言处理领域的新锐。同时，在预训练语言模型方面，基于 Transformer 的模型也呈现出了飞速发展的趋势，研究者们已经取得了可喜的进展，并且成果被广泛应用于各种实际场景之中。表 2.1 为目前已提出的主流预训练语言模型。本文将对 BERT 语言模型进行详细介绍，并借助该模型完成文本隐写任务。

表 2.1　基于 Transformer 的预训练语言模型

预训练语言模型	结构	预训练任务
GPT[42]、GPT-2[44]	Transformer 解码器	语言模型
BERT[41]	Transformer 编码器	掩码语言模型和下一句预测
RoBERTa[82]	Transformer 编码器	掩码语言模型
XLNet[83]	Transformer 编码器	排列组合语言模型
ELECTRA[84]	Transformer 编码器	掩码语言模型
UniLM[85]	Transformer 编码器	掩码语言模型和下一句预测
MASS[86]	Transformer 编码器和解码器	序列到序列掩码语言模型

2.3.4 BERT 语言模型

近年来，预训练语言模型如 BERT（Bidirectional Encoder Representations from Transformers）[41]，RoBERTa[82]（A Robustly Optimized BERT Pretraining Approach）等自然语言处理任务中均取得了先进的结果，被广泛应用于各种文本相关的应用场景。BERT 采用基于 Transformer 的结构，并使用基于字级别和词级别的嵌入方式进行输入表示。相对于传统的单向语言模型，BERT 提供双向语义建模，即在预测任务时考虑前后文信息来得到更准确的预测结果。此外，BERT 还引入了动态二进制掩码机制，来处理多任务学习场景中的数据泄漏问题。

其中，BERT 最核心的部分是预训练阶段，其对无标注数据进行预测能力的训练可以提高在特定下游任务上的表现。在这一阶段，BERT 使用大规模无标注语料库进行预训练，通过自监督学习的方式获取深层语言表示。具体而言，在两个预训练任务中，即掩码语言模型（Masked Language Model）和下一句预测（Next Sentence Prediction），BERT 这两个任务使用不同的学习策略和目标函数来捕捉单词和上下文之间的关系。

掩码语言模型是 BERT 的第一个预训练任务，其主要思想是：随机遮盖输入序列中的 15%单词，并让其去预测遮盖单词的原始形态。因此，模型能够从整个句子中理解被遮盖单词的含义和前后文的关联性。这样，BERT 就可以更好地感知上下文，处理真实场景下的自然语言数据，并为下游任务提供更加准确、可靠的输出结果。下一句预测是 BERT 的第二个预训练任务，它的目标是让模型能够理解两个句子之间的逻辑关系，判断它们是否紧密相关。这对于自然语言处理应用中的文本推理、问答和机器翻译等任务来说尤为重要，因为很多情况下需要考虑多个句子之间的语义联系。

从模型结构上来讲，BERT 的预训练模型是由多层 Transformer Encoder 模块组成，每层包含一个多头自注意力机制（Multi-Head Self-Attention Mechanism）和全连接前馈网络（Fully Connected Feed-Forward Network），该部分上小节中进行了介绍，这里不在赘述。这里主要介绍其输入部分，如图 2.11 展示了 BERT 的基本框架。BERT 的输入包括三个要素：词向量（Token Embeddings）、分割向量（Segment Embeddings）和位置向量（Positional Embeddings）。

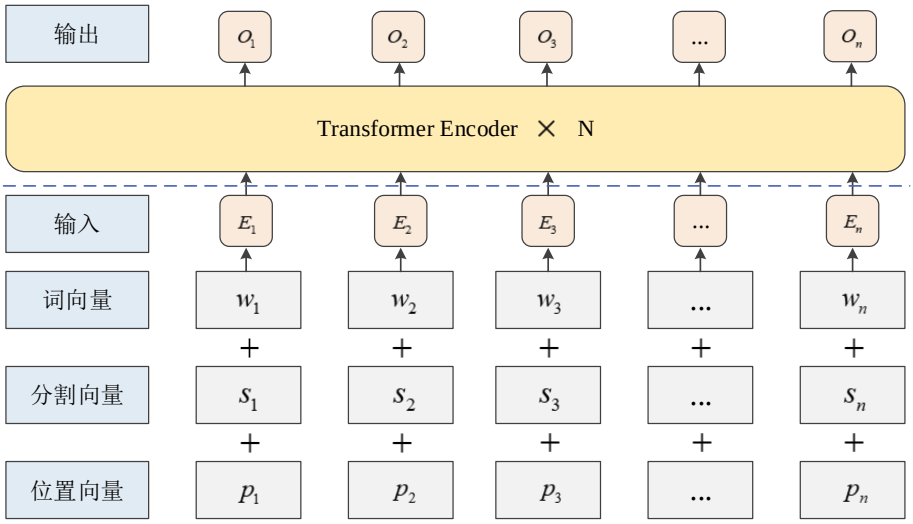


图 2.11　BERT 基本框架示意图

如图 2.11 所示，词向量对应输入序列中的每个单词，它将其表示成一个向量；分割向量则标识每个单词属于哪个句子；而位置向量则表示单词在序列中的位置。与 Transformer 相比，BERT 模型的输入多了一个分割向量部分。这是因为 BERT 的下一句预测任务，在训练时需要理解两个句子之间的关系。除此之外，BERT 还会在预训练阶段进行掩码语言模型任务，因此它需要标识掩码单词的位置并进行相应计算处理。因此，BERT 的输入结构是一种优化后的、更加适合自然语言表达的输入方式。

BERT 与传统的单向语言模型不同，在训练时采用了双向编码器结构，充分利用上下文中的双向信息。为了实现掩码语言模型的训练目标，BERT 将输入序列中的部分单词替换成特殊符号“[MASK]”，并要求模型基于上下文恢复出这些被替换的单词。在训练完成后，针对任何给定的被掩盖的单词，BERT 可以计算其在所有可能候选单词上的概率分布，从而实现高质量完形填空效果。这种双向编码方式相比传统的单向语言模型，更加适合处理自然语言表达，有效提升自然语言处理应用的性能。

2.4 本章小结

本章首先介绍了常用文本隐写的基本框架，并介绍了文本隐写方法，分为基于文本格式修改、基于文本内容修改和基于生成的文本隐写方法。接着介绍了文

本隐写评价指标，本文研究中主要利用不可感知性和嵌入容量作为文本隐写评价标准。然后，对文本隐写分析的基本概念和方法进行阐述，并介绍了文本隐写分析的评价指标。除此之外，根据语言模型的发展情况这一线索进行展开，并对BERT 语言模型和内部结构进行了着重介绍。

第三章　基于 BERT 和一致性编码的自回归文本隐写

3.1 引言

随着无线通信和社交网络服务的快速发展，许多人通过移动终端设备集成到社交网络中，即时分享日常生活的感受和媒体数据。这给文本隐写带来了极大的便利，将秘密信息传送给数据接收者而不引起监视者的怀疑。但是，由于文本本身的低冗余性，相比图像、音频和视频等数字媒体而言，文本隐写的研究更加复杂和具有挑战性。文本隐写一般可以分为两类：修改式文本隐写和生成式文本隐写。修改式文本隐写具有较高的语义隐藏性，但其有效载荷并不高。生成式文本隐写可以提供更高的嵌入有效负载，但其安全性不能很好地保证[47]。

最近，Ueoka 等人[87]提出了一种基于 BERT[41]的新型掩码语言模型来实现文本隐写，该方法简化了同义词典的构建规则，为修改式文本隐写方法提供了新的视角。虽然该方法生成的载密文本能够携带相对较高的有效载荷，但其在掩码位置的词语预测顺序是以并行的方式进行的。这意味着与掩码位置所对应的词具有高度的独立性。众所周知，文本中的每个单词都与文本的流畅性密切相关。如果这些单词具有高度的独立性，那么它将很容易被人类的感知系统识别出来，从而激励对手开发出先进的隐写分析器来降低安全性。因此，目前亟需一种更有效的策略来生成更高质量的载密文本。

本章提出了一种基于 BERT 和一致性编码的自回归文本隐写算法，该算法使用自回归策略根据秘密信息生成载密文本。本章主要贡献是，首先对从掩码语言模型中获得的候选词使用自回归策略，依次用可替换的词填充掩码位置，然后将它们输入到掩码语言模型中进行预测。同时，本章工作还改进了信息编码策略，针对给定的文本使用一致性编码来弥补块编码的缺点，这样就可以编码任意大小的候选词集，并利用概率分布进行信息隐藏。这模仿了母语使用者的单词选择机制，提高了句子的可读性和真实性。实验结果表明，与非自回归方法相比，该方法在保证安全性的同时提高了载密文本的流畅性，在一定程度上提高了嵌入的有效载荷，验证了本章工作的优越性。

3.2 相关工作

目前，Ueoka 等人[87]提出了一种新的基于 BERT 的掩码语言模型来实现文本隐写的方法，成功地将语言模型扩展到修改式文本隐写方法中。数学上，给定一个文本 $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)\subset V^n$，其中 x_i 是从一个大的词汇表 V 中抽取的第 i 个单词，n 是文本中的单词总数，该方法旨在生成这样一个载密文本 $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)\subset V^n$ 秘密信息 $b\subset\{0,1\}^L$ 可以从 $\boldsymbol{y}$ 中提取。数据嵌入过程可以描述如下。首先，根据密钥 k 从 $\boldsymbol{x}$ 中选择一定数量的单词位置作为掩码位置。然后，通过将非掩码位置的单词输入给掩码语言模型，可以确定每个掩码位置与预测概率相关联的候选单词列表。然后，根据 b 和引入的块编码策略，可以通过将 $\boldsymbol{x}$ 中掩码位置的单词替换为对应的候选列表中合适的单词来生成 $\boldsymbol{y}$。

为了成功地从 $\boldsymbol{y}$ 中提取 b，数据隐藏者和数据接收者应该共享密钥 k 和掩码语言模型 M。密钥 k 确保数据隐藏者和数据接收者能够识别相同的掩码位置。掩码语言模型 M 确保数据隐藏者和数据接收者为每个掩码位置的每个候选字获得相同的预测概率，以便数据隐藏者知道如何将秘密流编码为单词，并且数据接收者知道如何解码到对应的秘密流。

基于 BERT 的掩码语言模型来实现文本隐写的方法，与以往修改式文本隐写方法相比，该方法显著提高了嵌入的有效载荷，并在一定程度上提高了安全性。它也使修改式文本隐写方法和生成式文本隐写方法之间的关系更加接近。然而，该方法存在两个缺点，使得对手容易利用这两点来揭示秘密信息的存在。第一个问题是，该方法在掩码位置的词语预测是并行的。换句话说，不同掩码位置的单词预测是相互独立的，即预测特定位置的单词不会影响另一个位置的预测。然而，每个单词都与文本的流畅性密切相关。如果这些词具有高度的独立性，它将很容易被人类的感知系统识别，允许对手开发新的隐写分析器来降低安全性。因此，一种更有效的词预测策略的设计是迫切需要的。

另一方面，该方法中使用的信息编码策略，即块编码策略，只是简单地将每个掩码位置的候选词映射到具有固定长度的二进制流中。例如，对于一个特定的掩码位置，假设有 m 个单词的预测概率大于一个阈值 t_p，则可以确定满足 $2^l\le m$

的最大 l。上述方法使用概率最大的 2^l 个单词作为候选词，每个单词都恰好携带 l 个秘密位。这样，在数据嵌入过程中，使用与长度为 l 的秘密流相匹配的候选词来填充掩码位置。可以看出，虽然可以通过调整阈值 t_p 来控制长度 l，但它忽略了不同的单词具有不同的预测概率。例如，假设 $l = 2$，候选词表为“wonderful(0.6)”、“decent(0.2)”、“fine(0.1)”和“great(0.1)”，其中“wonderful(0.6)”意味着“wonderful”的预测概率是 0.6，上面的方法将每个单词映射到长度为 2 的二进制流，例如，单词可以分别映射为“00”、“01”、“10”和“11”。显然，该方法在数据嵌入过程中，所有填充掩码位置的候选词都具有相同的概率，即 $1/2^l$。然而，根据掩码语言模型，在数据嵌入过程中，尽可能选择一个预测概率高的单词，而不是随机选择一个单词，以获得更好的载密文本质量。这意味着设计一种更有效的信息编码策略是迫切需要的。

词预测策略和信息编码策略的选择会直接影响到文本隐写系统的安全性。词预测策略指的是在选择嵌入秘密信息时，是否考虑原始文本的含义、句法和语法结构等，并基于此进行预测。信息编码策略指的是如何对秘密信息进行编码并嵌入到载密文本中。因此，当前亟需开发一种新的文本隐写算法，这种算法应该能够在满足不可感知性和安全性需求的同时，尽可能确保隐写后的载密文本质量和可读性。更重要的是，为修改式文本隐写方法提供一个新的视角。

3.3 基于 BERT 和一致性编码的自回归文本隐写

3.3.1 总体框架

本章提出了一种基于 BERT 的自回归文本隐写方法，该算法采用自回归策略，根据需要嵌入的秘密信息生成载密文本。本章所提出的自回归文本隐写方法涉及三个参与者：数据隐藏者 Alice、攻击者 Wendy 和数据接收者 Bob。Alice 的目标是将二进制流的形式的秘密消息隐藏到载体文本中，具体而言，Alice 嵌入过程的框架如图 3.1 所示。由此产生的携带秘密信息的载密文本将通过一个不安全的通道发送给 Bob，比如互联网，该通道被对手 Wendy 监控，其目的是确定被传输的文本是否包含秘密信息。

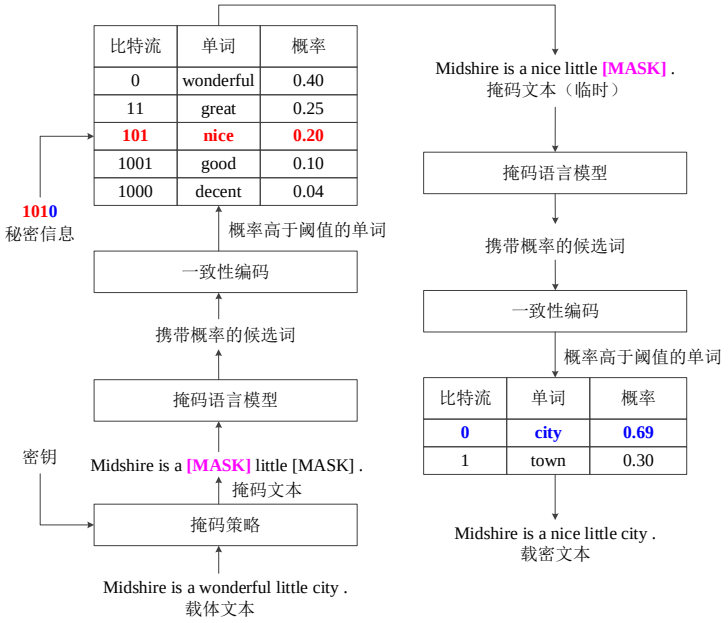


图 3.1　本章所提出方法的嵌入过程

具体来说，根据密钥和需要嵌入的秘密信息，Alice 首先在载体文本中确定一组掩码位置，对于要处理的每个掩码位置，将根据掩码语言模型、当前的临时文本和一致性编码技术共同确定一个单词，同时该单词需要与嵌入的秘密信息相匹配。当所有掩码位置都被处理完成后，Alice 可以生成载密文本并发送给 Bob，Bob 在载密文本上执行类似于 Alice 的操作，如图 3.2 所示。

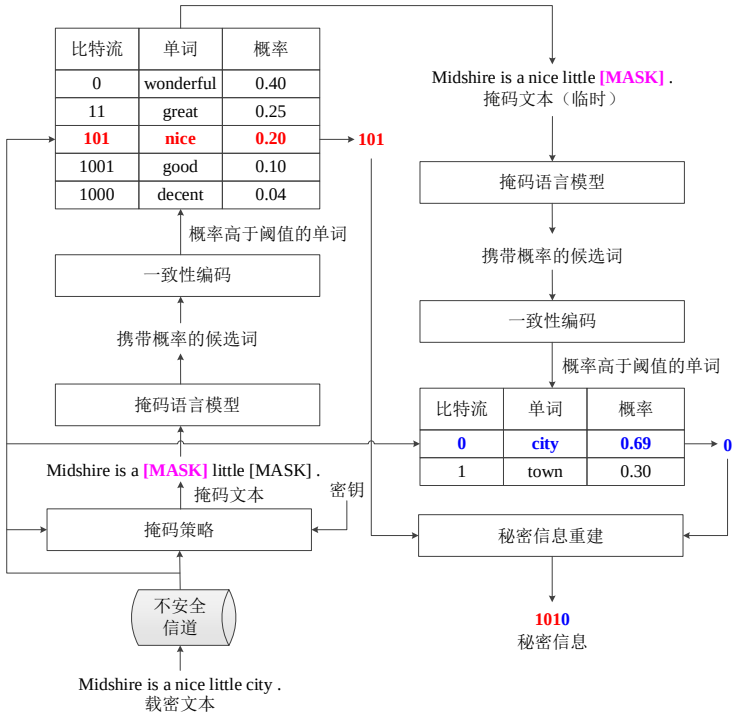


图 3.2　本章所提出方法的提取过程

为了确保数据提取过程能够成功执行，Alice 和 Bob 应该提前共享必要的边信息，包括密钥、控制候选词选择的阈值、掩码语言模型以及一致性编码策略。算法 3.1 展示了数据嵌入的伪代码。对于数据接收者，执行一个类似的过程从载密文本中重建秘密信息。

算法 3.1 数据嵌入过程的伪代码

输入：载体文本 $\boldsymbol{x}$，掩码语言模型 M，秘密信息流 b，参数 t_p，f 和密钥

输出：载密文本 $\boldsymbol{y}$

1．根据式(3.1)、f 和密钥确定 $\boldsymbol{x}_s$

2．for $i = 1,2,\ldots,n$ do

3．　if $x_i^{'}$ = [MASK] then

4．　　设置 $\boldsymbol{x}_{[i]} = \boldsymbol{x}_s$

5．　　根据 M 和 $\boldsymbol{x}_{[i]}$ 确定每个 $v_j \in V$ 的 p_j

6．　　用 t_p 确定候选词 $V' \subset V$

7．　　根据式(3.3)将 V' 中单词概率进行归一化

8．　　通过霍夫曼编码将 V' 中候选词映射到二进制流中

9．　　确定与 b 前缀相匹配的 x_i^*

10．　　更新 $\boldsymbol{x}_s$，将 $x_i^{'}$ 替换为 x_i^*

11．　　从 b 中删除已嵌入的前缀

12．　end if

13．end for

14．根据 $\boldsymbol{x}_s$ 确定 $\boldsymbol{y}$

15．返回 $\boldsymbol{y}$

3.3.2 掩码策略和掩码语言模型

掩码策略的目标是将载体文本中的一些单词替换为特殊的标记“[MASK]”。这些掩码位置将被由掩码语言模型基于上下文生成的词语填充。这些新的词语不仅适合上下文，而且还携带着秘密信息。我们可以自由地设计掩码策略。虽然掩

码策略可以精心制作，但它并不是本文的主要贡献。为了在实验中进行公平的比较，我们遵循了之前论文中引入的简单而有效的掩码策略。简单地说，一个整数 $f>0$ 可以看作是一个密钥，被用来控制掩码位置的数量。f 越高，这意味着掩码的词语越少，降低了嵌入容量，但增加了检测的难度。f 越高，这意味着掩码的词语越少，降低了嵌入容量，但增加了检测的难度。

基于修改的文本隐写的关键在于掩码语言模型，掩码语言模型引入了 BERT，BERT 的模型架构是基于 Vaswani 等人提出的 Transformer[40]的编码器。BERT 中提出了一种预训练目标：掩码语言模型（Masked Language Model），克服了单向进行训练的局限，实现了共同依赖于上下文进行预测。一般来说，可以根据特定的任务，如序列分类、词性标注或问答，使用预先训练好的 BERT 对下游任务做微调，可以大大降低设计自身架构的费用。然而，对于我们的任务不需要进行微调。与循环神经网络（Recurrent Neural Network, RNN）[37]一个接一个地看到单词和自回归预训练语言模型（Generative Pre-trained Transformer, GPT）[42]模糊未来标记不同，掩码语言模型由于其双向性提供了优越的预测性能。为此，我们使用预训练的 BERT 作为掩码语言模型。

设 M 为掩码语言模型，$I=\{i_1,i_2,\ldots,i_s\}$ 是由密钥确定的掩码位置的索引集。对于每个 $i\in I$，载体文本 $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ 中对应的单词 x_i 将被替换为“[MASK]”。为了在数学上表示它，对于任何一个 $S\subset I$，我们将 $\boldsymbol{x}_s$ 定义为：

$$\boldsymbol{x}_{\boldsymbol{S}}=\left(x_1^{'},x_2^{'},\ldots,x_n^{'}\right) \tag{3.1}$$

对于 $1\le i\le n$，有

$$x_i^{'}=\begin{cases}[\text{MASK}], & \text{if } i\in S\\ x_i, & \text{otherwise}\end{cases} \tag{3.2}$$

显然，$\boldsymbol{x}_{\boldsymbol{\Phi}}=\boldsymbol{x}$ 和 $\boldsymbol{x}_I$ 是将所有单词替换为“[MASK]”后的文本。为了在掩码位置中嵌入隐藏信息，本章工作对掩码位置进行了有序的处理。对于每个掩码位置，M 将根据之前生成的临时文本获得词汇表中每个单词预测概率。较高的概率意味着相应的单词更适合替换“[MASK]”以适应上下文。接下来，将确定每个掩码位置的临时掩码文本。

为了不失去泛化性，设$i_1 < i_2 < ... < i_s$。对于掩码索引为$i_j \in I$的临时文本通过如下过程进行确定。首先，对于$i_1 \in I$的临时文本，用$\boldsymbol{x}_{[i_1]}$表示。通过将$\boldsymbol{x}_{[i_1]}$输入语言模型 M 中，我们可以为每个$v_i \in V = \{v_1, v_2, ..., v_z\}$生成一个对应的预测概率$p_i \in [0,1]$，并且满足$\sum_{i=1}^{z} p_i = 1$。根据要嵌入的秘密位和本章所提出的一致性编码技术，我们选择一个单词$x_{i_1}^* \in V$来替换x中第i_1个位置的“[MASK]”。对于$i_2 \in I$，$\boldsymbol{x}_{[i_1]}$和$\boldsymbol{x}_{[i_2]}$之间唯一的区别在于，$\boldsymbol{x}_{[i_1]}$在第i_1个位置中使用“[MASK]”，但$\boldsymbol{x}_{[i_2]}$使用$x_{i_1}^*$。一般来说，$\boldsymbol{x}_{[i_{j-1}]}$和$\boldsymbol{x}_{[i_j]}$之间唯一的区别是，在第$i_{j-1}$个位置，$\boldsymbol{x}_{[i_{j-1}]}$使用“[MASK]”，但是$\boldsymbol{x}_{[i_j]}$使用的是$x_{i_{j-1}}^*$，对任何$2 \le j \le s$而言。

对图 3.1 进行解释，载体文本是“Midshire is a wonderful little city.”，存在两个掩码位置，从左到右进行编号。对于第一个掩码位置，要输入掩码语言模型的临时文本是“Midshire is a [MASK] little [MASK].”。在第一个掩码位置被载密单词“nice”替换后，输入掩码语言模型的第二个掩码位置的临时文本是“Midshire is a nice little [MASK].”。可以看出，当前掩码位置的临时文本生成的结果取决于先前的临时文本。该过程可以认为是一种自回归的方法。

3.3.3 一致性编码

一个最重要的问题是如何建立单词和秘密位之间的映射关系。为了解决这个问题，Ueoka 等人[87]使用一个阈值 $0 \leqslant t_p \leqslant 1$ 来收集一个预测概率高于 t_p 的候选词列表，然后将每个候选词映射到一个固定长度的二进制流中。换句话说，收集到的候选词具有相同的被选择的概率，将其填充当前掩码位置，这没有考虑到候选词的概率分布，不能很好地适应上下文。

如上所述，最好选择这样的词，即它不仅需要能够匹配嵌入的秘密位，而且该词语通过掩码语言模型获得的预测概率尽可能高。因为秘密位是均匀分布的，所以选择一个特定的单词作为输出的概率，实际上取决于该单词所携带的秘密信息的数量。换句话说，对于映射到具有相同长度的二进制流的任意两个单词，它

们通常具有相同的匹配概率。例如，在文献[87]中，如果所有的候选词都被映射到一个长度为$l>0$的二进制流，那么在2^l个候选词中选择任何一个作为当前输出的概率为$\frac{1}{2^l}$。它表示一旦映射关系完成，就不可以自由地选择这个单词了。因此，映射关系的构建本身必须考虑预测概率分布，这样对于预测概率高的词表，选择其中任一个单词的概率也很高。

因此，本章工作期望找到这样一种信息编码策略，即选择一个单词作为当前输出的概率与其由掩码语言模型得到的预测概率成正比。本章将这种信息编码策略视为一致性编码。与文献[87]相比，一致性编码有两个显著的优势：(1)候选词的数量可以是任意整数，而不是固定的 2 的幂次；(2)它基于词汇中每个单词的统计概率分布，考虑到单词的频率，使编码更有利于文本的正则化。这种机制模仿了母语使用者在选择单词时的优先级，并相应地提高了安全性。

本章工作可以自由地设计一致性编码。然而，信息论中许多经典的熵编码方法都可以用来实现一致性编码，如霍夫曼编码、算术编码和香农-范诺编码[40]。一致性编码的主要实现步骤在算法 3.1 的第 6-8 行，具体过程可以描述如下。为了简单起见，本章使用霍夫曼编码来提供现成的解决方案。形式上，为了紧凑性，对于每个特定的掩码位置，设$\{p_1,p_2,...,p_z\}$为词汇表$V=\{v_1,v_2,...,v_z\}$中单词的预测概率，其中$p_1\geq p_2\geq...\geq p_z$并且$\sum_{i=1}^{z}p_i=1$。首先收集所有预测概率高于预先确定的阈值 t_p 的单词，收集到的单词被视为要编码的候选单词。然后，通过对候选词的预测概率进行归一化，进一步应用霍夫曼编码将每个候选词映射到二进制比特流中。例如，让$V'=\{v_1,v_2,...,v_w\}\subset V,(w\leq z)$，包含候选词，归一化操作为：

$$p_i\leftarrow\frac{p_i}{\sum_{j=1}^{w}p_j},\forall 1\leq i\leq w \tag{3.3}$$

这使本章可以直接使用霍夫曼编码。众所周知，霍夫曼编码不会产生唯一的映射。例如，在图 3.1 中，“city”和“town”这两个词分别被映射到“0”和“1”。也有可能是“city”被映射到“1”，而“town”被映射到“0”。因此，不同的编码方案可能会产生不同的映射结果，如果数据隐藏者和数据接收者使用不同的密钥

来控制霍夫曼编码过程，从而导致秘密信息无法正确提取出来。为了避免这种情况的发生，数据隐藏者和数据接收者需要事先共享密钥。

3.4 实验结果与分析

本章实验使用了大规模的 CC-100 数据集[88]，该数据集由通过网络爬取到的高质量数据组成，并包含了 100 多种语言的单语数据。从 CC-100 数据集的英文部分随机抽取了 10,000 个文本，保证了其随机性和通用性，更好地证明了本章所提出方法的可靠性。需要注意的一件事是，每个载体文本的长度必须大于 20，这样才能嵌入足够的有效载荷。载体文本中嵌入一个随机生成的二进制秘密信息来生成载密文本。如第 3.3 节所述，本章实验使用 BERT 作为掩码语言模型。为了进行公平的比较，实验设置与文献[87]非常匹配，本章实验使用了谷歌的 $\text{BERT}_{\text{base, cased}}$ 模型和 transformers 包[40]的默认设置。此外，本章使用常用的指标困惑度（Perplexity, PPL）[60]来评估文本质量，一般来说，较低的 PPL 对应的文本质量越好，并通过隐写分析方法来衡量该方法的安全性。下面将对本章方法进行定性分析、定量分析和消融实验，并对实验结果进行分析。

3.4.1 定性分析

定性分析依靠人作为阅读者，根据他们的主观感受对文本进行分析，旨在更准确地评估文本的质量，从而更符合人类的感官需求。实验基于 BERT 和一致性编码的自回归文本隐写方法旨在将秘密信息嵌入到载体文本之中，并形成载密文本。在实验过程中，如表 3.1、表 3.2 和表 3.3 所示，展现了一些实验结果的直观的示例来评价本章方法所生成的载密文本的质量。此外，在第 3.2 节中提到与之前文献[87]所提出的方法之间的密切相关性。为了比较不同隐写方法的效果差异，本节进行了系统的对比实验。根据经验，将上节中定义的阈值默认设置为 0.02。

为了控制实验条件，本章使用嵌入参数 f 的变量来控制掩码位置的数量。当 f 值越高时，文本中的掩码位置就越少，这意味着文本修改较少可以提高文本的安全性。表 3.1、表 3.2 和表 3.3 分别展示了在 $f=2$、$f=3$ 和 $f=4$ 条件下，不同隐写方法生成的示例。在文本隐写中，有效载荷一般使用比特每个词（Bits Per

Word, BPW)进行描述。通过比较这些不同的数据表格，可以获得对整个实验的直观理解，并能够评估各种方法的优劣程度。

表 3.1　不同文本隐写方法的示例比较(f = 2)

隐写方法	载密文本	PPL	BPW
载体文本	Gerson and Walter contend there are multiple paths to solving a mathematical problem and students should be encouraged to chart their own way by exploring problems drawn from the real world.	-	-
文献[87]	Gerson and Walter contend there are multiple ways (2,1) to solving a particular (8,001) problem and students (4,10) should be encouraged to go (4,00) their own way by exploring (0,-) problems different (8,001) from the real one (4,10).	63.5261	0.42
本章方法	Gerson and Walter contend there are multiple approaches (3,1) to solving a given (7,001) problem and users (7,100) should be encouraged to find (4,000) their own way by solving (3,1) problems different (4,1) from the real one (5,000).	**56.4565**	**0.48**

表 3.2　不同文本隐写方法的示例比较(f = 3)

隐写方法	载密文本	PPL	BPW
载体文本	An unlucky accident befell my servant, Stevens, in falling from the coach and being dragged by the foot upon the pavement. Edition: current; Page: [186] He was in great danger, but happily was not essentially hurt.	-	-
文献[87]	An unlucky accident befell my friend (4,00), Stevens, in falling from the roof (8,101) and being dragged by the foot onto (2,1) the pavement. Edition: current ; Page: [186] He was in great shape (4,01), but happily was not essentially happy (4,11).	105.5982	0.27
本章方法	An unlucky accident befell my nephew (7,0010), Stevens, in falling from the window (9,1101) and being dragged by the foot across (3,11) the pavement. Edition: current ; Page: [186] He was in great debt (8,000), but happily was not essentially rich (7,1000).	**96.9807**	**0.46**

由于掩码策略总是跳过标点符号，因此标点符号永远不会携带数据。为简单起见，排除标点符号来确定文本中的单词总数，例如，在表 3.1 中，当f = 2 时，由文献[87]生成的载密文本的有效载荷为 13/31 = 0.42，而不是 13/32 = 0.41。在表 3.1 中，“$W(a, b)$”表示单词“W”被映射到一个二进制流 b，在数据嵌入过程

中，当前掩码位置的候选单词的数量为 a。例如，“ways (2,1)”的意思是在单词“ways”的位置上存在 2 个候选词，并且只包含一个秘密位“1”;“exploring (0,-)”意味着当前位置没有候选词可供选择，原始的词“exploring”被用于当前位置。

表 3.3　不同文本隐写方法的示例比较 ($f = 4$)

隐写方法	载密文本	PPL	BPW
载体文本	Accountability is a major key to BILSTEIN’s success. We hold our team members, line employees, frontline leaders, managers and company leaders responsible for thinking above the line in addressing both success and failure, taking direct responsibility for situations through personal steps to solve and overcome issues and not becoming a victim for individual or collective results.	-	-
文献[87]	Accountability is a major key (4,01) to BILSTEIN’s success. We hold our team leaders (2,1), line employees, frontline leaders, managers (4,00) and company leaders responsible for working (8,010) above the line in addressing both success and loss (4,01), taking direct responsibility for situations (0,-) through personal steps to solve and address (4,10) issues and not becoming a victim for individual (2,0) or collective results.	64.4565	0.23
本章方法	Accountability is a major key (5,011) to BILSTEIN’s success. We hold our team managers (3,00), line employees, frontline leaders, executives(5,010) and company leaders responsible for staying (11,01) above the line in addressing both success and failure (6,1), taking direct responsibility for moving (6,0010) through personal steps to solve and resolve (5,0) issues and not becoming a victim for mistakes (3,01) or collective results.	**57.0832**	**0.32**

3.4.2　定量分析

定量分析通常使用评估指标来衡量人类主观感受，并通过数值结果进行验证，以更客观地评估载密文本的质量。针对隐写分析的性能进行评估，对于每个实验，首先将上述 10,000 个自然文本及其载密文本分成三个不相干的子集，即训练集(60%)、验证集(10%)和测试集(30%)，然后用训练集和验证集对 $\text{BERT}_{\text{base, cased}}$ 模型进行微调。经过微调的模型被用来检测待测文本是否为载密文本。在测试集上用验证集精度最高的模型来评估准确性。实验中将本章所提出的方法与四种最

先进的方法，这四种方法分别来自文献[36]，[43]，[49]，[87]，进行比较，经过实验后得到了具体的比较结果，并整理在表格 3.4 中展示。

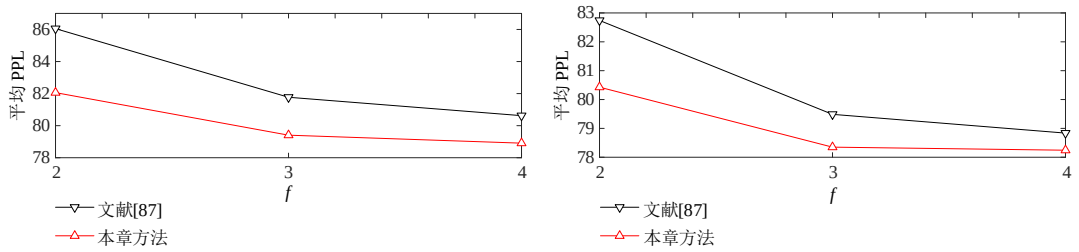
在实验中对参数进行了详细的设置，并致力于保证公平的比较结果。具体而言，实验中的参数设置描述如下，文献[87]和本章所提出的方法使用相同的参数设置，即 $f=3$ 和 $t_p=0.02$。τ 是文献[43]中引入的一个系统参数，实验中使用文献[43]中的推荐值，即 $\tau=0.7$。文献[49]的作者提出了两种方法，即固定长度编码和可变长度编码。实验中使用可变长度编码策略进行仿真，因为它具有更好的性能。对于文献[36]和[49]中对应的参数，截断长度 L 和块大小 B 均被设置为 2^2，达到公平比较的效果。

表 3.4　比较不同有效载荷的隐写方法的准确率

方法	参数	平均有效载荷	Acc
文献[87]	$f=3, t_p=0.02$	0.2479	0.5570
文献[43]	$\tau=0.7$	2.2934	0.8967
文献[49]	$L=2^2$	1.8072	0.9580
文献[36]	$B=2^2$	2.0000	0.9615
本章方法	$f=3, t_p=0.02$	0.2506	**0.5498**

如表 3.4 所示，针对不同的方法进行性能比较，在有效载荷方面，文献[36]、[43]和[49]方法的性能优于文献[87]方法和本章所提出的方法。原因是这三种方法实际上是基于生成的方法，而文献[87]中的方法和本章所提出的方法实际上是修改式文本隐写方法。生成式文本隐写方法通常能够提供高有效载荷，因为生成的每个单词都可以用来携带秘密信息。但在表 3.4 中，三种生成式方法的检测准确率（Accuracy, Acc）都很高，这意味着安全性没有得到满足。通过与文献[87]的比较，我们可以发现，本章所提出的方法不仅实现了较高的有效载荷，而且检测精度也较低，表明所提出的方法在有效载荷和安全性之间实现了更好的权衡。

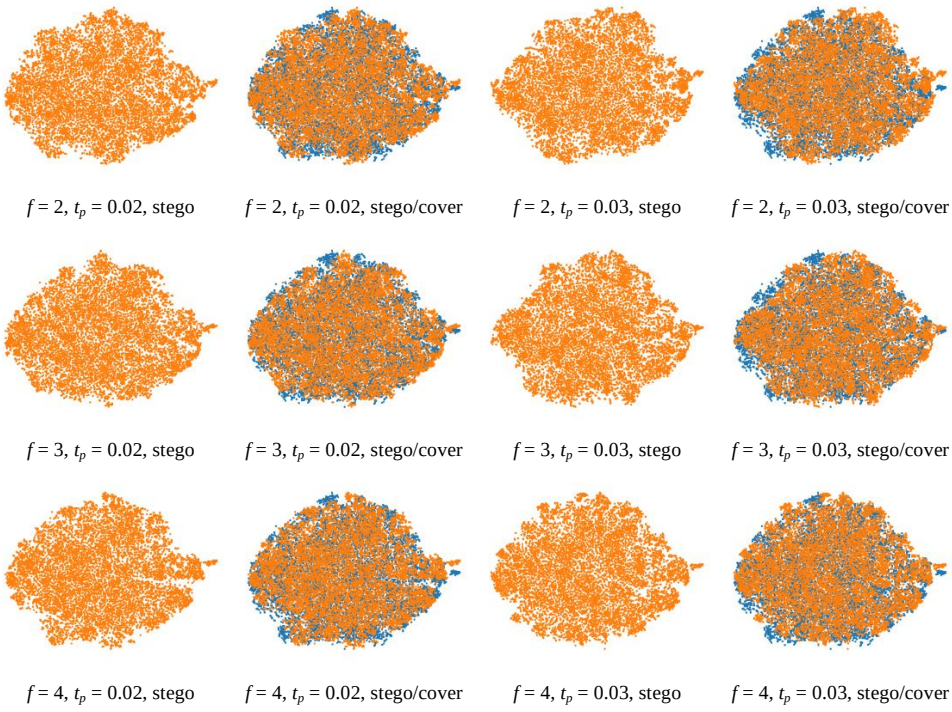
如图 3.2 展示了在不同参数条件下载密文本的平均 PPL，图 3.2(a)中使用 $t_p=0.02$，(b)中使用 $t_p=0.03$。可以推断出，本章方法比文献[87]生成的载密文本质量更好，文本流畅度高，生成的载密文本与自然文本更难区分，这揭示了本章所提出的方法具有良好的隐蔽性。



(a) t_p = 0.02　　(b) t_p = 0.03

图 3.2　在不同参数条件下载密文本的平均 PPL

为了进一步证明本章所提出的方法的优越性，采用了 t-SNE[89]可视化工具，将载密文本和自然文本的统计分布进行了对比。在该实验中，蓝点代表自然文本(cover)，橙色代表的是载密文本(stego)。从图 3.3 中可以看出，当有效载荷大小减少时，两种颜色的重叠点更多，表明载密文本更接近于自然文本的分布，因此说明确实具有更高的安全性，使其更难以被识别。



f = 2, t_p = 0.02, stego　　f = 2, t_p = 0.02, stego/cover　　f = 2, t_p = 0.03, stego　　f = 2, t_p = 0.03, stego/cover

f = 3, t_p = 0.02, stego　　f = 3, t_p = 0.02, stego/cover　　f = 3, t_p = 0.03, stego　　f = 3, t_p = 0.03, stego/cover

f = 4, t_p = 0.02, stego　　f = 4, t_p = 0.02, stego/cover　　f = 4, t_p = 0.03, stego　　f = 4, t_p = 0.03, stego/cover

图 3.3　利用 t-SNE[89]对载密文本和自然文本进行可视化分析

对于图 3.3 中所给出的不同参数，表 3.5 和表 3.6 分别展示了文献[87]和本章方法相应的平均有效载荷和准确率。本章实验在表 3.5 中使用 t_p = 0.02，在表 3.6 中使用 t_p = 0.03。从表 3.5 和表 3.6 可以推断出，在不同的参数设置下，本章所提出的方法在有效载荷和准确率方面均优于文献[87]的方法，从而验证了本章方法的优越性和有效性。

表 3.5　比较不同掩码间隔的隐写方法的检测精度（$t_p = 0.02$）

f	文献[87]		本章方法	
	有效载荷	Acc	有效载荷	Acc
2	0.3818	0.6107	**0.3906**	**0.6045**
3	0.2479	0.5570	**0.2506**	**0.5498**
4	0.1903	0.5403	**0.1937**	**0.5337**

表 3.6　比较不同掩码间隔的隐写方法的检测精度（$t_p = 0.03$）

f	文献[87]		本章方法	
	有效载荷	Acc	有效载荷	Acc
2	0.2916	0.6050	**0.3145**	**0.5968**
3	0.1919	0.5532	**0.2057**	**0.5465**
4	0.1474	0.5417	**0.1578**	**0.5407**

3.4.3　消融实验

本章方法通过使用自回归策略和一致性编码技术实现了卓越的性能。为了证明自回归策略和一致性编码技术对于提高安全性或有效载荷均起到了不可替代的作用。因此，本章进行了消融实验来验证这种假设。具体而言，在本节实验中通过单独移除自回归策略或者一致性编码技术的方式，对比其与完整模型的表现结果。文献[87]中的方法使用非自回归策略和块编码方法进行隐写。通过用自回归代替非自回归，实验中可以模拟出一种新的隐写方法，称为“仅自回归”方法。同样地，通过用一致性编码代替块编码，可以模拟出一种新的隐写方法，称为“仅一致性编码”方法。

如表 3.7 所示，在有效载荷方面，“仅一致性编码”和本章方法相互接近，但都高于文献[87]和“仅自回归”方法，这表明本章方法可以从一致性编码技术中受益，从而增加有效载荷。在隐写分析性能方面，“仅自回归”和本章方法有相似的表现，但都超过了文献[87]和“仅一致性编码”，这意味着自回归策略比非自回归策略的隐写算法具有更高的安全性。通过结合自回归策略和一致性编码技术，

本章方法的 PPL 值在大多数情况下是最低的。综上所述，可以得出结论，与相关工作相比，本章所提出的工作取得了最好的性能。

表 3.7　不同隐写策略之间的性能比较

f		2	3	4	2	3	4
t_p		0.02			0.03		
文献[87]	有效载荷	0.3818	0.2479	0.1903	0.2916	0.1919	0.1474
	PPL	86.0537	81.7708	80.6252	82.7407	79.4869	78.8377
	Acc	0.6107	0.5570	0.5403	0.6050	0.5532	0.5417
仅自回归	有效载荷	0.3818	0.2473	0.1897	0.2934	0.1923	0.1472
	PPL	84.8311	81.5283	80.3875	81.7111	79.3132	78.7971
	Acc	0.6048	**0.5487**	0.5398	0.6028	0.5512	0.5410
仅一致性编码	有效载荷	**0.3916**	**0.2532**	**0.1938**	0.3134	**0.2058**	**0.1578**
	PPL	83.1867	79.7653	78.9204	81.3915	78.5438	**78.2237**
	Acc	0.6123	0.5612	0.5423	0.6045	0.5520	0.5433
本章方法	有效载荷	0.3906	0.2506	0.1937	**0.3145**	0.2057	**0.1578**
	PPL	**82.0609**	**79.4073**	**78.9092**	**80.4290**	**78.3508**	78.2437
	Acc	**0.6045**	0.5498	**0.5337**	**0.5968**	**0.5465**	**0.5407**

此外，虽然本章方法使用 BERT 作为掩码语言模型，但实际上实验中可以自由地设置掩码语言模型。在本节实验中，采用 RoBERTa[82](A Robustly Optimized BERT Pretraining Approach)代替 BERT，以进一步证明本章方法的适用性和有效性。表 3.8 和表 3.9 中提供了不同参数下的实验结果，其中“RoBERTa”指的是用 RoBERTa 替换文献[87]中使用的语言模型 BERT。“RoBERTa+Proposed”则意味着用 RoBERTa 代替本章所提出的方法中使用的语言模型 BERT。本章实验使用了不同的参数，在表 3.8 中使用 $t_p = 0.02$，在表 3.9 中使用 $t_p = 0.03$。从表 3.8 和表 3.9 可以推断出，本章所提出的方法提高了文本隐写性能，这也验证了其适用性和优越性。实验结果表明，与其他现有方法相比，本章提出的方法在各项指标上均取得了较为显著的改进，并且最终获得了最佳的效果。这些结果证明了该方法的有效性和可靠性。

表 3.8　RoBERTa 作为掩码语言模型，不同掩码间隔下隐写方法的检测精度(t_p = 0.02)

f	RoBERTa		RoBERTa+Proposed	
	有效载荷	PPL	有效载荷	PPL
2	0.7649	94.2367	**0.7966**	**74.0341**
3	0.4454	72.6491	**0.4498**	**67.4493**
4	0.3235	69.2259	**0.3239**	**65.9934**

表 3.9　RoBERTa 作为掩码语言模型，不同掩码间隔下隐写方法的检测精度(t_p = 0.03)

f	RoBERTa		RoBERTa+Proposed	
	有效载荷	PPL	有效载荷	PPL
2	0.6359	85.7642	**0.6498**	**72.0499**
3	0.3658	68.7636	**0.3808**	**65.6212**
4	0.2642	66.6071	**0.2767**	**64.7378**

3.5 本章小结

本章提出了一种基于 BERT 模型的自回归文本隐写算法，并引入了一致性编码。通过利用自回归词预测策略，生成的单词能更好地融入上下文，从而提高了生成的载密文本的可读性和真实性。同时，一致性编码技术扩大了文本隐写的候选词数量，并利用了候选词的统计概率分布，这样不仅提高了有效载荷，而且使载密文本看起来更加自然。实验表明，与相关工作相比，本章所提出的方法取得了最佳性能，验证了其优越性和适用性。

从实用的角度来看，随着文本在社交网络上的广泛使用，文本隐写将变得越来越流行。通过社交网络平台传递载密文本，隐写行为很容易被大量的普通社交活动所掩盖。更重要的是，一旦数据接收者观察到载密文本，他可以保持沉默，在不采取任何可疑互动的情况下提取秘密信息，这就掩盖了数据接收者的真实身份。尽管近年来文献中报道的文本隐写方法越来越多，但如何在有效载荷、文本质量和安全性之间实现良好的权衡仍然是一个具有挑战性的问题。

一方面，尽管修改式文本隐写方法很好地保持了文本的语义质量，但可嵌入

的有效载荷却很低。另一方面，尽管生成式文本隐写方法允许在载体文本中嵌入较高的有效载荷，但需要付出语义信息不可控和安全性降低的高昂代价。在本章中，通过利用语言模型进行单词预测，并利用概率分布进行信息编码，与修改式文本隐写方法相比，本章所提出的工作在保证安全的同时提高了载密文本的流畅性，并与生成式文本隐写方法相比提高了安全性。有效载荷也在一定程度上增加，这使得修改式文本隐写方法更接近于生成式文本隐写方法。后续将探索更有效的策略来增加可嵌入的有效载荷，以便在修改式文本隐写方法和生成式文本隐写方法之间建立一座桥梁。

第四章　基于掩码语言模型的可逆文本信息隐藏

4.1 引言

信息隐藏作为一种有效的秘密通信手段，使我们能够利用数字媒体的冗余性，将秘密信息隐蔽地嵌入数字媒体中。新生成的含有秘密信息的媒体不会引入明显的伪影，因此媒体的使用不会受到影响，并且可以实现秘密信息传输和版权保护等目的。尽管许多信息隐藏方法都能成功地提取出秘密信息，但被改变的媒体内容会出现永久性失真，这表明原始媒体内容不能被完美地恢复。这对于一些敏感的应用场景，也就是需要接收方完全恢复原始媒体的场景，这是不可取的。

因此，近年来人们对可逆信息隐藏[2]进行了深入的研究来处理上述问题。与传统的信息隐藏算法相比，可逆信息隐藏算法允许数据接收者在从嵌入载体中提取秘密信息后，能够不失真地恢复原始载体内容。也就是说，通过信道进行秘密传输后，当用户被授权获得相同的密钥时，可以从含有秘密信息的媒体中无损地恢复原始载体和秘密信息。

目前，可逆信息隐藏算法已经被广泛地应用于数字图像和视频媒体中[90, 91]，但在文本中的可逆信息隐藏成果却非常少。原因是文本具有高度编码的特性，几乎没有冗余，可供嵌入的信息位数比图像和视频少。因此，有必要为文本中的可逆信息隐藏设计新的方案。在众多媒体类型中，文本是人类最常使用的信息载体。此外，文本还具有高鲁棒性的特性，通过信道传输的变化很小，而且不容易受到噪声的干扰。这意味着可逆文本信息隐藏具有良好的应用前景，这促使本文作者对文本中的可逆信息隐藏进行研究。

如图 4.1 所示，文本中的可逆信息隐藏可以简单地描述如下。数据隐藏者需要设计这样一种数据嵌入算法，他接收一个载体文本、一个密钥和秘密数据作为输入，然后输出一个带有秘密数据的载密文本。由此产生的载密文本将通过一个不安全的渠道发送给数据接收者。根据密钥，数据接收者能够从载密文本中提取秘密数据，并无误地重建载体文本。通过这种方式，文本中的可逆信息隐藏得以实现。可逆文本信息隐藏的一个最重要的要求是，载密文本应该看起来是正常的，

也就是说，载密文本不应引入令攻击者产生怀疑的痕迹，从而被发现秘密信息的存在。因此，这需要保证载密文本的流畅性和语义质量，同时，载密文本需要能够抵御统计检测工具。

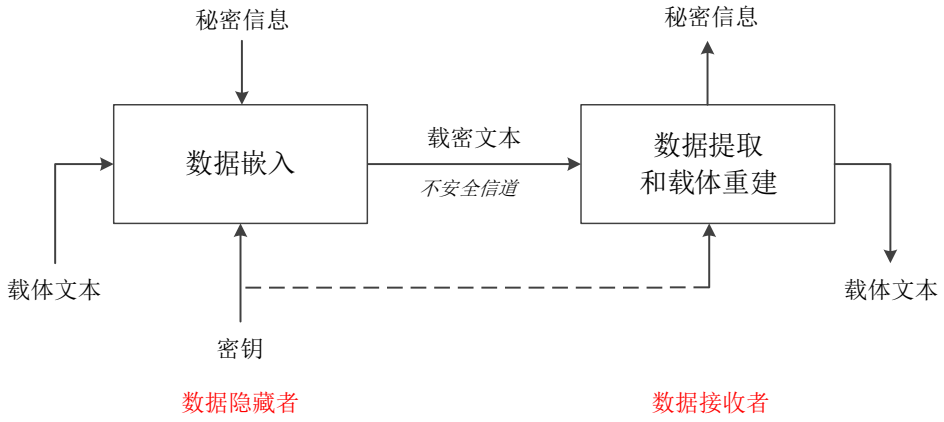


图 4.1　可逆文本信息隐藏的框架

在文本中实现可逆信息隐藏的一个简单方法是以无损的方式改变文本格式。例如，可以调整两个相邻的单词或相邻行之间的空白，以容纳秘密信息[92,93]。通过从载密文本的格式中提取秘密位，可以进一步恢复原始文本，因为原始文本的字是不变的。另一个简单的想法是改变载体文本中的字符，这使得秘密位可以被嵌入，但可能导致不正确的字。为了恢复载体文本，这些由于信息嵌入引起的不正确的字应该被纠正。然而，不正常的文本格式和不正确的字很容易引起对手的怀疑，从而降低了载密文本的不可感知性。

人们可以将最初为数字图像设计的可逆嵌入策略扩展到文本。例如，Liu 等人[94]将载体文本中的一些单词转换成整数，然后通过在图像中广泛使用的整数变换和差分扩展将秘密信息嵌入到单词中。尽管这种方法确保了可逆性，但由于上溢/下溢的问题，纯有效载荷的大小相当低。此外，虽然它可以被改进，从而在一定程度上增加有效载荷的大小，但新的问题，如大量边信息会出现[95]。这表明，将应用于数字图像的可逆方法扩展到文本并不容易。

随着深度学习[96]和自然语言处理的快速发展，应用于文本的主流信息隐藏方法使用训练好的语言模型来达成信息嵌入的目的[47,61]。然而，这些方法不能确保其可逆性。最近，Chang[97]提出了一种利用索引扩展将秘密比特流可逆地嵌入到载体文本的方法。在该方法中，一个训练有素的语言模型被用来预测要嵌入的位

置的单词，这使得数据隐藏者能够收集所有候选单词的预测概率表。这些单词可以根据预测概率与一个索引相关联。因此，通过确定与要嵌入的当前比特相匹配的索引，可以选择相应单词作为输出。尽管该方法在数据嵌入能力和语义失真之间表现出良好的权衡，但它要求数据隐藏者和数据接收者事先共享预训练的语言模型，这在实际应用中存在较大的局限性。一方面，语言模型包含了许多参数，意味着数据隐藏者和数据接收者之间共享的边信息太多，这在实践中是不可取的。另一方面，数据隐藏者可能不想与数据接收者分享训练有素的语言模型，因为训练好的语言模型可以被视为数据隐藏者的数字资产。此外，使用语言模型进行数据提取意味着较高的计算复杂性，这对于实际应用是不利的。

因此，为文本开发更有效的可逆信息隐藏方案的需求是极为迫切的，本章提出了一个基于掩码语言模型的可逆文本信息隐藏的算法。本章所提出的算法的主要思想是生成这样一个载密文本，通过收集一些特定位置的词来重建载体文本，并对其他位置的词进行处理以提取秘密信息。因此，秘密信息和原始载体文本可以成功地隐藏在载密文本中，并可以从载密文本中完美地提取出来。实验结果表明，本章所提出的工作确保了载体文本和秘密信息的可逆性，并提供了良好的文本质量和安全性，表明了本章方法的优越性。

4.2 基于掩码语言模型的可逆文本信息隐藏

4.2.1 总体框架

本章提出了一种基于掩码语言模型的可逆文本信息隐藏方法，该方法同样遵循图 4.1 所示的基本框架。其基本思路是利用载体文本作为“语义控制密钥”来引导生成携带秘密信息的载密文本。与传统使用修改给定的载体文本的方法不同，本章方法更加注重文本的语义和结构完整性，可以有效地降低信息嵌入和数据提取时被检测到的风险。图 4.2 显示了本章方法的总体框架。从图 4.2 中可以推断出，所提出的框架中有五个重要部分，即语义初始化、语义控制、数据嵌入、数据提取和载体重建。下面，本章将详细介绍本章所提出的可逆文本信息隐藏方法的五个核心部分。

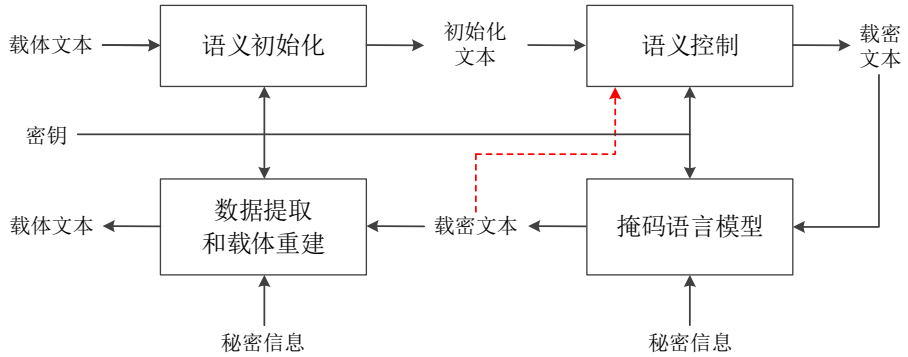


图 4.2　本章方法的总体框架

4.2.2　语义初始化

给定一个载体文本和一个位置密钥，语义初始化的目的是生成一个初始化文本，将其用于后续的数据嵌入过程。具体来说，让 $\boldsymbol{c}=(c_1,c_2,...,c_n)$ 表示载体文本，其中 c_i，$1\le i\le n$ 是从一个非常大的词汇表 V 中采样的第 i 个单词。位置密钥 $P=\{p_1,p_2,...,p_n\}$ 是一个满足 $1\le p_1<p_2<...<p_n$ 的整数序列。语义初始化的目标是生成初始化文本 $\boldsymbol{u}=\{u_1,u_2,...,u_m\}$，其需要满足 $p_n\le m$ 和公式(4.1)。

$$u_i^{'}=\begin{cases} c_i, & \text{if } i=p_j\in p \\ [\text{MASK}], & \text{otherwise} \end{cases} \tag{4.1}$$

其中，“[MASK]”表示一个特殊的标记。例如，假设 $\boldsymbol{c}=(I,do,.)$，$P=\{1,7,12\}$ 和 $m=12$，$\boldsymbol{u}$ 可以确定为{ ***I***, [MASK], [MASK], [MASK], [MASK], [MASK], ***do***, [MASK], [MASK], [MASK], [MASK], . }，图 4.3 展示了语义初始化的过程。

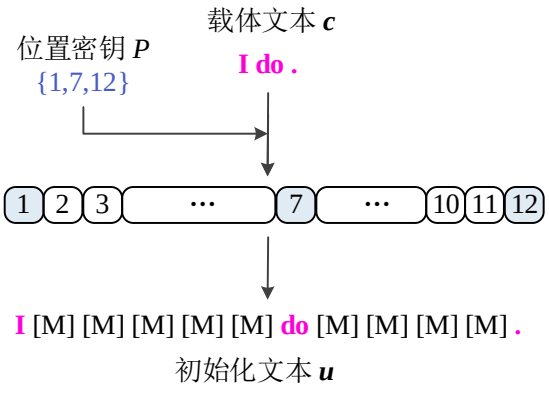


图 4.3　语义初始化过程的示例

4.2.3 语义控制

语义控制的目的是生成一个载密文本，并将其输入到后续的掩码语言模型模块中。语义控制的算法是自由设计的，意味对于载密文本的操作，除了与载体文本相对应的单词不被改变之外，其余没有严格的限制。本章方法简单地将载密文本设置为初始化文本或者临时文本，这是由于后续的嵌入方法是一个迭代过程的原因，参考下一小节。在图 4.2 中，从载密文本到语义控制块的红色虚线是迭代过程的具体体现，图 4.4 对该迭代过程展示了一个更为具体的示例。

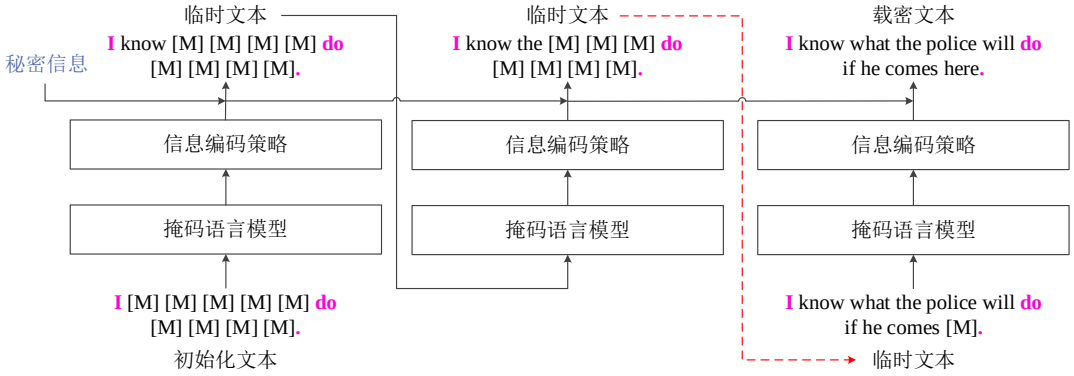


图 4.4 语义控制过程的示例

4.2.4 数据嵌入

掩码语言建模(Masked Language Modeling)任务的目的是生成一个携带秘密信息的载密文本，这是一个迭代过程。在每次迭代中，本章使用预先训练好的语言模型，根据掩码文本和嵌入的秘密数据生成一个临时文本。掩码文本和临时文本之间的区别是，掩码文本的一个掩码位置被一个单词取代后得到临时文本。除非所有的掩码位置都被处理，否则临时文本将被用来替换下一次迭代的掩码文本。研究表明，在主流的自然语言处理任务中，掩码语言建模任务是文本生成任务中常用的策略。在本章中，掩码语言模型被用于生成携带秘密信息的文本。因此，本章中掩码语言模型的作用实际上等同于数据嵌入的过程。

数据嵌入的伪代码显示在算法 4.1 中。该过程可以看出，通过预测每个掩码位置的单词，可以生成最终的载密文本。然而，如算法 4.1 的第 5 行所示，本章应该使用信息编码方法来确定特定掩码位置的单词，这可以是自由设计的。信息

编码的目的是从候选词列表中选择一个候选词来替换特殊标记“[MASK]”，同时与要嵌入的秘密信息相匹配。由于这不是本章的主要创新所在，为了简单起见，我们使用现有的信息编码方法。本章实验将评估各种信息编码，并在 4.3.2 小节中详细描述了四种有代表性的信息编码策略的技术细节。

算法 4.1 数据嵌入过程的伪代码

输入：载体文本 $\boldsymbol{u}$，语言模型 M，秘密信息流 b

输出：载密文本 $\boldsymbol{s}$

1．初始化 $\boldsymbol{s} = \boldsymbol{u}$

2．for $i = 1,2,\ldots,m$ do

3．　if s_i = [MASK] then

4．　　用 M 和 $\boldsymbol{s}$ 生成一个候选词表，每个单词都与 s_i 的预测概率相匹配

　　//注：$\boldsymbol{s}$ 在这里被称为掩码文本

5．　　根据信息编码方法确定与 b 前缀相匹配的候选词 w

6．　　利用 $s_i = w$ 更新 $\boldsymbol{s}$ //注：$\boldsymbol{s}$ 在这里被称为临时标记文本

7．　　从 b 中删除已嵌入的前缀

8．　end if

9．end for

10．返回 $\boldsymbol{s}$

4.2.5 数据提取

数据提取的目标是从载密文本中提取秘密信息 b，数据提取的过程与数据嵌入的过程类似。它取决于算法 4.1 第 5 行中所示的信息编码方法。数据接收者对载密文本进行操作，并使用位置密钥 $P = \{1,7,12\}$ 和 m = 12 获得与数据隐藏者相同的初始化文本 $\boldsymbol{u}$，经过同样的步骤，就可以完全恢复秘密信息。以桶编码为例，V_b 中的词被映射到秘密位 $b \in \{0,1\}$ 中，因此嵌入的秘密位可以从词汇的一个子集中确定。具体来说，如果单词和秘密位之间的映射关系是基于候选词的预测概率，那么信息编码和信息解码都由预训练的语言模型 M。控制。这表明，预训练的语言模型 M。应该在数据隐藏者和数据接收者之间共享，因为数据接收者需要使用

语言模型来重构词和秘密比特之间的映射关系。否则，数据隐藏者就没有必要与数据接收者共享 M。显然，从实际使用的角度来看，只有数据隐藏者持有语言模型是比较理想的。然而，从嵌入性能的角度来看，共享语言模型可能更好。在实验中发现前者取得了更有竞争力的性能，这意味着不共享语言模型的方法对于可逆信息隐藏是一个合适的策略。

4.2.6 载体重建

对于数据接收者而言，直接根据共享的位置密钥 P 从载密文本 $\boldsymbol{s}$ 中提取对应位置的单词进行组合即可重建载体文本。重建原始的载体文本的方法是很简单的。它可以被描述为

$$c_i = s_{p_i}, \forall 1 \le i \le n \tag{4.2}$$

其中，位置密钥 $P = \{p_1, p_2, ..., p_n\}$ 应该提前在数据隐藏者和数据接收者之间共享。

该载体重建方法，也就是将载密文本中的数据提取出来后仍然能够无损恢复载体的过程。这一载体重建过程所需的时间复杂度和空间复杂度极低，这意味着它可以快速地完成载体重建过程，并且占用较少的硬件资源。此外，该方法不要求数据隐藏者和数据接收者之间共享语言模型。这使得双方共享的边信息较少，从而将潜在的信息泄露风险最小化。因此，本章所提出的方法非常适用于数据接收方行为受限制以及即时通信场景中。

4.3 实验结果与分析

本节将提供实验结果和性能评估的分析，以验证本章所提出方法的可行性和适用性。为此，本章进行了一系列实验，并利用相关指标进行评价。首先建立了一个实验环境，本节给出了实验设置和评估指标。同时，为了公平地比较，实验中使用不同的信息编码策略进行展开，从而本节对四种代表性的信息编码策略进行了介绍。最后，通过这些实验，本节得到了详尽的实验结果和多个性能指标，如准确率、有效载荷和困惑度等。综合来看，这些实验结果和性能评估有助于加深对该方法的理解，并为进一步开展相关研究提供有益的参考。

4.3.1 实验设置和评估指标

本章使用 Python 和 PyTorch 进行实验。对于语言模型，本章选用 Hugging Face 开源[98]的 transformers 工具包中预训练好的 $\text{BERT}_{\text{base, uncased}}$ 模型。本章实验中使用了 BookCorpus[99]这一基准数据集，它是一个在自然语言处理任务中被广泛使用的大型文本语料库，由约 18,000 本主题丰富、蕴含不同细粒度语义信息的书籍构成。我们随机选择 10,000 个载体文本用于可逆文本信息隐藏。需要注意的是，每个载体文本的长度，即 n，需要控制在一个合理的范围内，这样可以降低计算的复杂性，同时保证载密文本的质量。而且，为了简单起见，我们假设载密文本的长度是载体文本的倍数，即 $n \mid m$。m 可以自适应地增加，成为一个可以整除 n 的较大的整数，为了使得载密文本以一个停止符结束。

为了评估本章所提方法产生的载密文本的质量，使用了 PPL 值来进行衡量。PPL 计算的是句子中每个单词的平均对数概率。一般来说，PPL 值越低，生成的文本就越自然，载密文本也就越安全。本章使用 GPT-2（Generative Pre-trained Transformer 2）[44]计算所有载密文本的平均 PPL 作为评判标准。为了评估安全性，我们使用 10,000 个自然文本和 10,000 个具有相同长度范围和相似语法结构的载密文本，用文献[100]中的方法进行检测分析。其中，两个常见的指标：准确性（Accuracy, Acc）和 F1 值（F1-score, F1），被用于衡量抵抗检测分析的能力。另一个指标是数据嵌入的有效载荷，它被定义为每个单词携带的平均比特数（Bits Per Word, BPW），该指标尽可能高。

4.3.2 信息编码策略

为了公平地比较，本章使用不同的信息编码策略进行实验。四种有代表性的信息编码策略，即块编码[27]、霍夫曼编码[59]、自适应分组编码[101]（Adaptive Dynamic Grouping, ADG）和桶编码[36]在本章实验中被评估比较。为了自成一体，本节将简要介绍了它们的技术细节。

（1）块编码

给定一组候选词，块编码方法根据预测概率从该组中选出的 2^k 个单词分配

到从 0 到2^k-1的k位二进制码中。例如，对于一个特定的掩码位置，通过将词汇表V中的所有候选词按照预测概率从高到低排序，每个最前面的2^k词都可以与一个长度为k的二进制代码相关联。我们需要使用一个阈值t_p来控制V中可用词的数量，也就是说，对于一个给定的掩码位置，只有V中预测概率大于t_p的词可以用来携带秘密信息。另一方面，数据接收者应持有t_p和掩码语言模型，以便他能从载密文本中提取嵌入的数据。

（2）霍夫曼编码

霍夫曼编码是可变字长编码的一种编码方式，在基于生成的隐写算法中被多次使用到，这种编码方式根据单词出现的概率来构造平均长度最短的码字，这样可以有效地减少冗余编码。若选择一个单词作为当前输出的概率与其由掩码语言模型得到的预测概率成正比，则这种信息编码策略视为一致性编码。霍夫曼编码本质上就是一种一致性编码技术。上述块编码方法为单词分配固定长度的编码，而霍夫曼编码则根据预测概率为单词分配长度不确定的编码。与块编码一样，霍夫曼编码需要t_p和语言模型来嵌入数据和提取数据。

（3）自适应分组编码

自适应分组编码的目标是将词汇中的所有单词分成一定数量的组，以便每个组代表一个唯一的秘密二进制流。为了嵌入秘密数据，对于一个掩码位置，从相应的组中抽出一个单词作为输出。对于数据提取，数据接收者应该有语言模型和词汇表，并且知道分组算法。为了保持一致性，t_p被用来控制可用词的数量。

（4）桶编码

桶编码提前将词汇中的所有单词映射成二进制流。在数据嵌入过程中，对于一个掩码位置，在与要嵌入的秘密数据相匹配的单词列表中，具有最大预测概率的单词被作为输出。例如，词汇表V可以分为两个不相交的子集V_0和V_1，其中$|V_0|\approx|V_1|$。V_b中的单词被映射到秘密位$b\in\{0,1\}$。在数据嵌入过程中，如果秘密位是b，选择V_b中预测概率最大的词作为输出。显然，数据接收者在不知道该词的预测概率的情况下，很容易提取秘密数据。换句话说，与上述方法相比，数据隐藏者和数据接收者不需要共享语言模型，这大大减少了边信息。默认情况下，在实验中，我们把V分成两个子集，这意味着，每个词承载了一个比特的秘密信息。

需要指出的是，数据隐藏者和数据接受者不需要存储V_b，因为通过应用哈希函数可以实现单词和秘密位之间的映射关系，这对数据提取非常方便。

4.3.3 实验结果和性能评估

首先提供一些具体的例子来验证本章所提方法的可行性。如表 4.1 所示，我们有$\boldsymbol{c}=(I,do,.)$，$P=\{1,7,12\}$和$m=4n=12$。无论嵌入的有效载荷的长度如何，都可以推断出生成的载密文本具有令人满意的质量。在表 4.1 中，“Block”、“Huffman”、“ADG”和“Bins”分别对应块编码、霍夫曼编码、自适应分组编码和桶编码。“Top-1”表示没有嵌入任何信息，即每个掩码位置总是被具有最大预测概率的词所填充。

表 4.1 不同载密文本的示例，其中 $m/n=4$

编码方式	t_p	载密文本
Block	0.02	***I** was going to try to **do** the same for myself.*
	0.03	***I** have no way to really **do** the same without him.*
	0.04	***I** do this, but i do not **do** this now.*
Huffman	0.02	***I** know what the police will **do** if he comes here.*
	0.03	***I** do it. you always **do** it. I know.*
	0.04	***I** always **do** it for a living.*
ADG	0.02	***I** have no one willing to **do** a little crazy thing.*
	0.03	***I** did the same thing you **do** for the same reason.*
	0.04	***I** did. but you can **do** that for her too.*
Bins	-	***I** know the way they can **do** it to me now.*
Top-1	-	***I** know what i have to **do** to keep her safe.*

为了衡量载体文本的流畅性，根据经验，我们在实验中简单地将 n 限制在[4,8]的范围内。如表 4.2 所示，确定了不同信息编码策略生成的载密文本的平均 PPL 值。随着信息编码策略中的 t_p 不断增加，PPL 值逐渐减小，这是因为较高的 t_p 意味着针对文本进行较少的修改，从而导致生成的载密文本的失真较少，更接近于自然文本，因此，得到载体文本更流畅，PPL 值更小。此外，当 m/n 比值增大时，PPL 值也呈逐渐下降的趋势，这是因为较长的文本往往会因为获得更多的语境使得文本可读性更佳。

表 4.2　比较不同信息编码策略的平均 PPL 值，其中 $n \in [4,8]$

信息编码策略	t_p	$m/n=3$	$m/n=4$	$m/n=5$
Block	0.02	223.8311	157.5531	128.8145
	0.03	202.9565	134.4043	103.8719
	0.04	187.6439	118.5401	91.9489
Huffman	0.02	202.6945	138.0253	108.4690
	0.03	189.8209	126.0925	97.6912
	0.04	180.3138	116.3862	89.3254
ADG	0.02	260.7759	189.2484	157.2128
	0.03	216.1935	153.3190	125.8440
	0.04	199.3847	134.2199	104.6727
Bins	-	274.5535	179.6919	134.2634

表 4.3 展示了不同的信息编码策略导致的平均有效载荷大小，从实验结果中可以发现，不同的信息编码策略导致了不同的有效载荷大小。同时，当 t_p 增加时，平均有效载荷的大小将下降。因为 t_p 越高，表明用于携带秘密数据的单词越少，从而导致有效载荷大小降低。

表 4.3　比较不同信息编码策略的平均有效载荷，其中 $n \in [4,8]$

信息编码策略	t_p	$m/n=3$	$m/n=4$	$m/n=5$
Block	0.02	1.2113	1.3822	1.4691
	0.03	0.9538	1.0794	1.1402
	0.04	0.7865	0.8839	0.9315
Huffman	0.02	1.2269	1.3946	1.4792
	0.03	1.0166	1.1464	1.2154
	0.04	0.8576	0.9620	1.0168
ADG	0.02	0.4734	0.5335	0.5507
	0.03	0.3561	0.3977	0.4113
	0.04	0.2742	0.3015	0.3104
Bins	-	0.6507	0.7259	0.7660

针对表 4.3 中的编码策略进行比较，对于桶编码，尽管在大多数情况下有效载荷的大小小于块编码和霍夫曼编码，但有效载荷大小可以通过将词汇表 V 划分为更多不相干的子集来增加。例如，通过将 V 分成四个互不相干的子集，每个词可以用来携带两个比特信息，使得有效载荷大小增加一倍，从而在嵌入有效载荷方面优于其他方法。此外，更高的 m / n 意味着更多的掩码位置被用于数据嵌入，相应地导致获得更高的有效载荷，这在表 4.3 中已经被验证。

为了进一步衡量载密文本的安全性，如表 4.4 所示，实验评估了载密文本抗检测分析的能力，选取 10,000 个自然文本和 10,000 个载密文本组成样本集，并且按照 6:1:3 的比例分为训练集、验证集和测试集，最终由验证精度最高的模型对测试集进行评估，得到实验结果。

表 4.4　比较不同信息编码策略的检测准确率，其中 $n \in [4,8]$

信息编码	m / n	3		4		5	
策略	t_p	Acc	F1	Acc	F1	Acc	F1
Block	0.02	0.8917	0.8934	0.9213	0.9211	0.9418	0.9416
	0.03	0.9040	0.9030	0.9247	0.9254	0.9473	0.9469
	0.04	0.8938	0.8925	0.9232	0.9216	0.9425	0.9438
Huffman	0.02	0.8992	0.8987	0.9195	0.9202	0.9435	0.9441
	0.03	0.9020	0.9028	0.9250	0.9260	0.9535	0.9540
	0.04	0.9002	0.9018	0.9248	0.9245	0.9497	0.9495
ADG	0.02	0.9255	0.9267	0.9475	0.9478	0.9537	0.9542
	0.03	0.9127	0.9143	0.9397	0.9407	0.9540	0.9538
	0.04	0.9087	0.9069	0.9398	0.9402	0.9573	0.9568
Bins	-	0.9098	0.9094	0.9305	0.9290	0.9515	0.9521

从表 4.4 可以推断出，在大多数情况下，当 t_p 增加时，检测准确率(Acc)和 F1 值(F1)逐渐下降。这是合理的，因为较高的 t_p 对应于较低的嵌入有效载荷大小，导致自然文本和载密文本之间的统计差异较低，使得检测准确率和 F1 值较低。当 m / n 增加时，检测准确率和 F1 值逐渐增加，这是由于较长的文本更容易暴露出统计特性异常。然而，尽管不同的信息编码策略会导致不同的检测分析性

能，但在大多数情况下，不同策略之间的性能差异是接近的。从实际使用的角度来看，使用不需要数据隐藏者和数据接收者分享较多边信息的信息编码策略将是更可取的方式。换句话说，从实用性方面，根据表 4.4 可得，针对可逆文本信息隐藏的方法，使用桶编码是更为合理的。

4.4 本章小结

本章提出了一个基于掩码语言模型的可逆文本信息隐藏的算法，它完全不同于以往通过将数字图像的可逆嵌入策略扩展到文本中的方法。在本章所提出的框架中，通过根据位置密钥将载体文本分配到被掩码文本中，在掩码文本的掩码位置上填充单词以嵌入秘密数据。实验结果表明，秘密数据可以从载密文本中准确无误地被提取出来，并用预先共享的位置密钥可以完美地恢复原始载体文本。通过对参数的微调，可以达到足够的有效载荷。此外，生成的载密文本具有良好的质量和令人满意的抗检测分析的能力。

在未来，我们将进一步提高嵌入性能，我们希望这个框架能够激发更多先进的工作。在未来的研究中，我们将致力于进一步提高载密文本的嵌入性能和实用性。为此，我们将考虑应用更为灵活、智能化的信息编码策略。同时，我们也将探索运用先进的信息隐藏算法，以更好地保障载密文本的不可感知性。此外，我们希望通过该框架的研究，促进该相关的领域的发展和创新，激发更多针对该问题的研究工作。通过持续不断的技术改进和优化，我们希望这个框架可以在信息隐藏领域发挥越来越重要的作用。

第五章　结论与展望

5.1 结论

文本是信息传递的重要方式之一，本文研究的内容是语言模型驱动的文本隐写技术，利用语言模型实现在文本中嵌入秘密信息的过程。论文详细梳理了国内外的文本隐写技术的研究现状，并着重探讨了如何生成质量高，并携带秘密信息的文本这一关键问题。当前，深度学习技术在自然语言处理中得到广泛应用，最新主流算法采用语言模型来生成载密文本，相比传统技术有更大的有效载荷。但是，在保证安全性方面仍需加强。在此背景下，本文研究语言模型驱动的文本隐写，主要工作总结如下：

本文提出了一种自回归文本隐写算法，它基于 BERT 预训练模型和一致性编码的技术。与主流的非自回归方法相比，该工作在嵌入容量与系统安全性之间实现了更好的平衡。该方法使用掩码语言模型得到预测词及其概率分布，并利用一致性编码技术可以对任意数量的候选词集进行编码，并嵌入秘密信息。对于能够嵌入的掩码位置，该方法采用自回归的方式预先填入候选词进行预测，以增强上下文的联系，增加了先验条件，进而提升了文本的质量。经过一系列实验测试，相比于非自回归的文本隐写方法，该算法不仅能够保证安全性，还可以提高载密文本的流畅性，同时，在嵌入容量方面也取得了一定程度的提升。

本文提出了一种基于掩码语言模型的全新框架，可以实现可逆的文本信息隐藏。由于现有文本信息隐藏方法能够完美地提取秘密信息，但载体文本会出现一定程度的失真，这些算法具有不可逆性。这就意味着，如果没有提前共享足够多的边信息，则无法完美地恢复载体文本。因此，该算法主要基于掩码语言模型，数据隐藏者通过选择某些位置的词生成载密文本，数据接收者则利用与数据隐藏者类似的操作来提取嵌入的秘密信息，并实现对载体文本的重建。实验结果表明，该方法在保证了载体文本的安全性的同时，也能够完整地还原载体文本。而且，在所提出的方法中，数据隐藏者和数据接收者无需共享过多的边信息，这一定程度上降低了部署成本，增加了该方法在实际应用中的可能性。

5.2 展望

本文引入语言模型实现文本隐写，语言模型可以通过对给定的文本进行分析，从中学习出该自然语言的规律和特点，并结合秘密信息生成“合适”的载密文本。该类方法一定程度上提高了嵌入容量，同时，可以有效避开一些传统的文本隐写检测手段。但是，由于技术的限制，目前基于语言模型实现文本隐写还存在许多问题，例如文本嵌入容量比较受限。此外，这类方法也可能会增加文本的噪声和冗余信息，从而影响文本的可读性和真实性。因此，未来的研究工作可以从以下几个方面展开：

（1） 尽管本文提出的方法生成的载密文本在评价指标和统计分析检测方面已经取得了一定进展，但在语义上与人类所写的文章相比仍存在差距。因此，除了保证信息嵌入容量和不可感知性等性能外，如何进一步提高载密文本的语义相似度，更加准确、真实地呈现原始文本的意思，是未来研究所需要关注的重要议题之一。这需要不断拓展基于深度学习和自然语言处理技术的实践，以更好地满足实际信息隐藏需求。同时，在提高载密文本的语义相似度时，也要充分考虑隐蔽性和安全性的影响，确保密文不易被发现和识别。

（2）当前，针对载体文本语言为英文的隐写技术相对比较成熟，并且得到较广泛的应用。未来可以考虑针对其他语言的文本开展相关研究，并结合人工智能、自然语言处理等方面的技术实现更为复杂的嵌入和提取。例如，针对非英文文本进行文本隐写，需要考虑其他语言文字的不同特征，如词汇、语法结构等方面的差异，并针对具体语言设计算法模型。同时，可以结合机器翻译技术来实现多语言文本的嵌入和提取，从而拓宽文本隐写在多语言环境下的应用场景。

因此，文本隐写将会是一个充满机遇和挑战的领域，期待看到更多优秀的研究成果面世，并为信息安全保障做出巨大贡献。

参考文献

[1] BASH B A, GHEORGHE A H, PATEL M, et al. Quantum-secure covert communication on bosonic channels[J]. Nature Communications, 2015, 6(1): 8626.

[2] WU H, SHI Y, WANG H, et al. Separable reversible data hiding for encrypted palette images with color partitioning and flipping verification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 27(8): 1620-1631.

[3] TAN J, LIAO X, LIU J, et al. Channel attention image steganography with generative adversarial networks[J]. IEEE Transactions on Network Science and Engineering, 2021, 9(2): 888-903.

[4] WANG J, ZHANG L Y, CHEN J, et al. Compressed sensing based selective encryption with data hiding capability[J]. IEEE Transactions on Industrial Informatics, 2019, 15(12): 6560-6571.

[5] HASSABALLAH M, HAMEED M A, AWAD A I, et al. A novel image steganography method for industrial internet of things security[J]. IEEE Transactions on Industrial Informatics, 2021, 17(11): 7743-7751.

[6] WU H, LIU G, YAO Y, et al. Watermarking neural networks with watermarked images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2591-2601.

[7] WANG Y, YE J, WU H. Generating watermarked speech adversarial examples[C]//Proceedings of the ACM Turing Award Celebration Conference, July 30-August 1, 2021, Hefei, China. New York: ACM, 2021: 254-260.

[8] CHEN Y, WANG H, WU H, et al. Adaptive video data hiding through cost assignment and STCs[J]. IEEE Transactions on Dependable and Secure Computing, 2019, 18(3): 1320-1335.

[9] FENG B, LU W, SUN W. Secure binary image steganography based on minimizing the distortion on the texture[J]. IEEE Transactions on Information Forensics and Security, 2014, 10(2): 243-255.

[10] REKIK S, GUERCHI D, SELOUANI S A, et al. Speech steganography using wavelet and Fourier transforms[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2012, 2012: 1-14.

[11] KHOSLA S, PARAMJEET K. Secure Data Hiding Technique Using Video Steganography and Watermarking-A Review[J]. International Journal of Computer Applications, 2014, 95(20): 0975-8887.

[12] TASKIRAN C M, TOPKARA U, TOPKARA M, et al. Attacks on lexical natural language steganography systems[C]//Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents VIII, January 15, 2006, San Jose, USA. Bellingham: SPIE, 2006: 97-105.

[13] LOW S H, MAXEMCHUK N F, BRASSIL J T, et al. Document marking and identification using both line and word shifting[C]//Proceedings of the Conference on Computer Communications, Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies, April 2-6, 1995, Boston, USA. Piscataway: IEEE, 1995: 853-860.

[14] BRASSIL J T, LOW S, MAXEMCHUK N F. Copyright protection for the electronic distribution of text documents[J]. Proceedings of the IEEE, 1999, 87(7): 1181-1196.

[15] CHOO H G, KIM W Y. Data-hiding capacity improvement for text watermarking using space coding method[J]. Digital Watermarking, Springer Berlin Heidelberg, 2004, 2939: 593-599.

[16] 康慧娴, 易标, 吴汉舟. 文本隐写及隐写分析综述[J]. 应用科学学报, 2021, 39(6): 923-938.

[17] BRASSIL J T, LOW S, Maxemchuk N F, et al. Electronic marking and identification techniques to discourage document copying[J]. IEEE Journal on Selected Areas in Communications, 1995, 13(8): 1495-1504.

[18] BHAYA W, RAHMA A M S, AL-NASRAWI D. Text steganography based on font type in ms-word documents[J]. Journal of Computational Science, 2013, 9(7): 898-904.

[19] MAHATO S, YADAV D K, KHAN D A. A novel approach to text steganography using font size of invisible space characters in microsoft word document[C]//Proceedings of the Intelligent Computing, Networking, and Informatics: Proceedings of the International Conference

on Advanced Computing, Networking, and Informatics, June 12-14, 2013, Raipur, India. Berlin: Springer, 2014: 1047-1054.

[20] TANG X, CHEN M. Design and implementation of information hiding system based on RGB[C]//Proceedings of the 2013 3rd International Conference on Consumer Electronics, Communications and Networks, November 20-22, 2013, Xianning, China. Piscataway: IEEE, 2013: 217-220.

[21] LIU T Y, TSAI W H. A new steganographic method for data hiding in microsoft word documents by a change tracking technique[J]. IEEE Transactions on Information Forensics and Security, 2007, 2(1): 24-30.

[22] LIU Y, SUN X, LIU Y, et al. Mimic-ppt: Mimicking-based steganography for microsoft power point document[J]. Journal of Information Technology, 2008, 7: 654-660.

[23] ZHONG Z, XU G. Digital watermarking algorithm based on structure of PDF document[J]. Journal of Computer Applications, 2012, 32(10): 2776.

[24] POPA R. An analysis of steganographic techniques[J]. The Politehnica University of Timisoara, Faculty of Automatics and Computers, Department of Computer Science and Software Engineering, 1998, 65: 1-65.

[25] BOLSHAKOV I A. A method of linguistic steganography based on collocationally-verified synonymy[C]//Proceedings of the Information Hiding: 6th International Workshop, May 23-25, 2004, Toronto, Canada. Berlin: Springer, 2004: 180-191.

[26] 甘灿, 孙星明, 刘玉玲, 等. 一种改进的基于同义词替换的中文文本信息隐藏方法[J]. 东南大学学报: 自然科学版, 2007, 37(A01): 137-140.

[27] CHANG C Y, CLARK S. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method[J]. Computational Linguistics, 2014, 40(2): 403-448.

[28] TOPKARA M, TOPKARA U, ATALLAH M J. Words are not enough: sentence level natural language watermarking[C]//Proceedings of the 4th ACM International Workshop on Contents Protection and Security, Berlin, Germany, November 4-8, 2006. New York: ACM, 2006: 37-46.

[29] MURPHY B, VOGEL C. The syntax of concealment: reliable methods for plain text information hiding[C]//Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents IX, January 28, 2007, San Jose, USA. Bellingham: SPIE, 2007: 351-362.

[30] CHANG C Y, CLARK S. The secret's in the word order: Text-to-text generation for linguistic steganography[C]//Proceedings of the International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 8-15, 2012, Mumbai, India. Mumbai: Indian Institute of Technology Bombay, 2012: 511-528.

[31] DAI W, YU Y, DAI Y, et al. Text steganography system using Markov chain source model and DES algorithm[J]. Journal of Software, 2010, 5(7): 785-792.

[32] MORALDO H H. An approach for text steganography based on Markov chains[J]. arXiv preprint arXiv:1409.0915, 2014.

[33] LUO Y, HUANG Y, LI F, et al. Text steganography based on ci-poetry generation using Markov chain model[J]. KSII Transactions on Internet and Information Systems, 2016, 10(9): 4568-4584.

[34] NIU Y, WEN J, ZHONG P, et al. A hybrid R-BILSTM-C neural network based text steganalysis[J]. IEEE Signal Processing Letters, 2019, 26(12): 1907-1911.

[35] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[J]. arXiv preprint arXiv:1409.2329, 2014.

[36] FANG T, JAGGI M, ARGYRAKI K. Generating steganographic text with LSTMs[J]. arXiv preprint arXiv:1705.10742, 2017.

[37] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[38] YANG Z, ZHANG S, HU Y, et al. VAE-Stega: Linguistic steganography based on variational auto-encoder[J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 880-895.

[39] KANG H, WU H, ZHANG X. Generative text steganography based on LSTM network and attention mechanism with keywords[J]. Electronic Imaging, 2020, 2020(4): 291-1-291-8.

[40] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 1-11.

[41] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2-7, 2019, Minneapolis, USA. Stroudsburg: ACL, 2019: 4171-4186.

[42] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[43] ZIEGLER Z M, DENG Y, RUSH A M. Neural linguistic steganography[J]. arXiv preprint arXiv:1909.01496, 2019.

[44] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[45] YI B, WU H, FENG G, et al. ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling[J]. IEEE Signal Processing Letters, 2022, 29: 687-691.

[46] LUO Y, HUANG Y. Text steganography with high embedding rate: Using recurrent neural networks to generate chinese classic poetry[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, June 20-22, 2017, Philadelphia, USA. New York: ACM, 2017: 99-104.

[47] GUO Y, WU H, ZHANG X. Steganographic visual story with mutual-perceived joint attention[J]. EURASIP Journal on Image and Video Processing, 2021, 2021(1): 1-14.

[48] WU H, YI B, DING F, et al. Linguistic steganalysis with graph neural networks[J]. IEEE Signal Processing Letters, 2021, 28: 558-562.

[49] YANG Z, GUO X, CHEN Z, et al. RNN-stega: Linguistic steganography based on recurrent neural networks[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1280-1295.

[50] KRISHNAN R B, THANDRA P K, BABA M S. An overview of text steganography[C]//Proceedings of the 2017 Fourth International Conference on Signal Processing, Communication and Networking, March 16-18, 2017, Chennai, India. Piscataway: IEEE, 2017: 1-6.

[51] SIMMONS G J. The prisoners' problem and the subliminal channel[C]//Proceedings of the Advances in Cryptology, August 21-24, 1983, Santa Barbara, California. Boston: Springer, 1983: 51-67.

[52] LOW S H, MAXEMCHUK N F, LAPONE A M. Document identification for copyright protection using centroid detection[J]. IEEE Transactions on Communications, 1998, 46(3): 372-383.

[53] XIANG L, GUO G, YU J, et al. A convolutional neural network-based linguistic steganalysis for synonym substitution steganography[J]. Mathematical Biosciences and Engineering, 2020, 17(2): 1041-1058.

[54] KERMANIDIS K L, MAGKOS E. Empirical paraphrasing of modern Greek text in two phases: an application to steganography[C]//Proceedings of the Computational Linguistics and Intelligent Text Processing: 10th International Conference, March 1-7, 2009, Mexico City, Mexico. Berlin: Springer, 2009: 535-546.

[55] KERMANIDIS K L. Hiding secret information by automatically paraphrasing modern Greek text with minimal resources[C]//Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence, October 27-29, 2010, Arras, France. Piscataway: IEEE, 2010: 379-380.

[56] MILLER G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[57] CAN G, XINGMING S, YULING L, et al. An improved steganographic algorithm based on synonymy substitution for chinese text[J]. Journal of Southeast University (Natural Science Edition), 2007, 37(S1): 137-140.

[58] LLOYD S P. Binary block coding[J]. Bell System Technical Journal, 1957, 36(2): 517-535.

[59] HUFFMAN D A. A method for the construction of minimum-redundancy codes[J]. Proceedings of the IRE, 1952, 40(9): 1098-1101.

[60] BROWN P F, PIETRA V J D, MERCER R L, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.

[61] SHEN J, JI H, HAN J. Near-imperceptible neural linguistic steganography via self-adjusting arithmetic coding[J]. arXiv preprint arXiv:2010.00677, 2020.

[62] KANG H, WU H, ZHANG X. Generative text steganography based on LSTM network and attention mechanism with keywords[J]. Electronic Imaging, 2020, 2020(4): 291-1-291-8.

[63] 梁小萍, 何军辉, 李健乾, 等. 隐写分析——原理, 现状与展望[J]. 中山大学学报: 自然科学版, 2004, 43(6): 93-96.

[64] LI L, HUANG L, ZHAO X, et al. A statistical attack on a kind of word-shift text-steganography[C]//Proceedings of the 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, August 15-17, 2008, Harbin, China. Piscataway: IEEE, 2008: 1503-1507.

[65] SUI X, LUO H, ZHU Z. A steganalysis method based on the distribution of first letters of words[C]// Proceedings of the Second International Conference on Intelligent Information Hiding and Multimedia, December 18-20, 2006, Pasadena, USA. Piscataway: IEEE, 2006: 369-372.

[66] XIANG L, SUN X, LUO G, et al. Research on steganalysis for text steganography based on font format[C]//Proceedings of the Third International Symposium on Information Assurance and Security, August 29-31, 2007, Manchester, United Kingdom. Piscataway: IEEE, 2007: 490-495.

[67] XIANG L, SUN X, LUO G, et al. Linguistic steganalysis using the features derived from synonym frequency[J]. Multimedia Tools and Applications, 2014, 71: 1893-1911.

[68] MENG P, HANG L, YANG W, et al. Linguistic steganography detection algorithm using statistical language model[C]//Proceedings of the International Conference on Information Technology and Computer Science, August 8-11, 2009, Beijing, China. Piscataway: IEEE, 2009, 2: 540-543.

[69] CHEN Z, HUANG L, YU Z, et al. A statistical algorithm for linguistic steganography detection based on distribution of words[C]//Proceedings of the Third International Conference on Availability, Reliability and Security, March 4-7, 2008, Barcelona , Spain. Piscataway: IEEE, 2008: 558-563.

[70] ZHAO X, CHEN Z, HUANG L, et al. Effective linguistic steganography detection[C]//Proceedings of the 8th International Conference on Computer and Information Technology Workshops, 2008 8-11, July, Sydney, Australia. Piscataway: IEEE, 2008: 224-229.

[71] YANG Z, WEI N, SHENG J, et al. TS-CNN: Text steganalysis from semantic space based on convolutional neural network[J]. arXiv preprint arXiv:1810.08136, 2018.

[72] LI H, JIN S. Text steganalysis based on capsule network with dynamic routing[J]. IETE Technical Review, 2021, 38(1): 72-81.

[73] NIU Y, WEN J, ZHONG P, et al. A hybrid R-BILSTM-C neural network based text steganalysis[J]. IEEE Signal Processing Letters, 2019, 26(12): 1907-1911.

[74] BAO Y J, YANG H, YANG Z L, et al. Text steganalysis with attentional LSTM-CNN[C]//Proceedings of the 5th International Conference on Computer and Communication Systems, May 15-18, 2020, Shanghai, China. Piscataway: IEEE, 2020: 138-142.

[75] XU Y, ZHAO T, ZHONG P. Small-scale linguistic steganalysis for multi-concealed scenarios[J]. IEEE Signal Processing Letters, 2021, 29: 130-134.

[76] KUMARI R, SRIVASTAVA S K. Machine learning: A review on binary classification[J]. International Journal of Computer Applications, 2017, 160(7): 1-5.

[77] BELLEGARDA J R. Statistical language model adaptation: review and perspectives[J]. Speech Communication, 2004, 42(1): 93-108.

[78] BROWN P F, DELLA PIETRA V J, DESOUZA P V, et al. Class-based n-gram models of natural language[J]. Computational Linguistics, 1992, 18(4): 467-480.

[79] MONTAVON G, SAMEK W, MÜLLER K R. Methods for interpreting and understanding deep neural networks[J]. Digital Signal Processing, 2018, 73: 1-15.

[80] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model[C]//Proceedings of the 11th Annual Conference of the International Speech Communication Association, September 26-30, 2010, Makuhari, Japan. New York: ISCA, 2010: 1045-1048.

[81] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[82] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[83] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in Neural Information Processing Systems, 2019, 32: 1-11.

[84] CLARK K, LUONG M T, LE Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.

[85] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[J]. Advances in Neural Information Processing Systems, 2019, 32: 1-13.

[86] SONG K, TAN X, QIN T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[87] UEOKA H, MURAWAKI Y, KUROHASHI S. Frustratingly easy edit-based linguistic steganography with a masked language model[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 6-11, 2021, Online. Stroudsburg: ACL, 2021: 5486-5492.

[88] WENZEK G, LACHAUX M A, CONNEAU A, et al. CCNet: Extracting high quality monolingual datasets from web crawl data[J]. arXiv preprint arXiv:1911.00359, 2019.

[89] BELKINA A C, CICCOLELLA C O, ANNO R, et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets[J]. Nature Communications, 2019, 10(1): 5415.

[90] ZHANG X. Reversible data hiding in encrypted image[J]. IEEE Signal Processing Letters, 2011, 18(4): 255-258.

[91] CHEN Y, WANG H, TANG X, et al. A novel two-dimensional reversible data hiding method with high embedding capacity in H. 264/advanced video coding[J]. International Journal of Distributed Sensor Networks, 2020, 16(3): 1550147720911001.

[92] KUMAR R, MALIK A, SINGH S, et al. A space based reversible high capacity text steganography scheme using font type and style[C]//Proceedings of the International Conference on Computing, Communication and Automation. Piscataway: IEEE, 2016: 1090-1094.

[93] ALATTAR A M, ALATTAR O M. Watermarking electronic text documents containing justified paragraphs and irregular line spacing[C]//Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents VI, January 18-22, 2004, San Jose, USA. Bellingham: SPIE, 2004: 685-695.

[94] LIU Z, SUN X, LIU Y, et al. Invertible transform-based reversible text watermarking[J]. Information Technology Journal, 2010, 9(6): 1190-1195.

[95] FEI W, TANG X. Reversible text watermarking algorithm using prediction-error expansion method[C]//Proceedings of the International Conference on Computer, Networks and Communication Engineering, May 23-24, 2013, Beijing, China. Paris: Atlantis Press, 2013: 401-405.

[96] LECUN Y, BENGIO Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[97] CHANG C C. Reversible linguistic steganography with bayesian masked language modeling[J]. IEEE Transactions on Computational Social Systems, 2022.

[98] WOLF T, DEBUT L, SANH V, et al. Huggingface's transformers: State-of-the-art natural language processing[J]. arXiv preprint arXiv:1910.03771, 2019.

[99] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. Piscataway: IEEE Computer Society, 2015: 19-27.

[100] YANG Z, WANG K, LI J, et al. TS-RNN: Text steganalysis based on recurrent neural networks[J]. IEEE Signal Processing Letters, 2019, 26(12): 1743-1747.

[101] ZHANG S, YANG Z, YANG J, et al. Provably secure generative linguistic steganography[J]. arXiv preprint arXiv:2106.02011, 2021.

作者在攻读硕士学位期间公开发表的论文

[1] **ZHENG X**, WU H. Autoregressive linguistic steganography based on BERT and consistency coding[J]. Security and Communication Networks, 2022, 2022: 1-11. **(SCI: 000805159500006, EI: 20222312192455)**

[2] **ZHENG X**, FANG Y, WU H. General Framework for Reversible Data Hiding in Texts Based on Masked Language Modeling[C]//Proceedings of the 24th IEEE International Workshop on Multimedia Signal Processing, September 26-28, 2022, Shanghai, China. Piscataway: IEEE, 2022: 1-6. **(EI: 20225013233833)**

[3] WU H, YANG T, **ZHENG X**, FANG Y. Linguistic steganography and linguistic steganalysis, In book: Adversarial Multimedia Forensics, to appear, Springer, 2023.

作者在攻读硕士学位期间所参与的项目

[1] 国家自然科学基金青年项目“社交网络多用户协同的行为隐写”(项目编号：61902235).

致　　谢

岁月不居，时节如流。不知不觉中研究生三年的学习时间已经临近尾声。三年里所发生过的点点滴滴仍历历在目恍如隔日，回想起这段时光，仍记得刚开始时的兴奋与迷茫，慢慢地开始有了自己的目标并且在追求过程中愈发坚定，一路上成长了许多。在即将毕业之际，我谨向身边经常帮助和支持我的老师、同学们还有我的朋友家人们致以我最真诚的谢意和最美好的祝福。

首先我要感谢我的导师吴汉舟老师对我的谆谆教诲，吴老师严谨的教学态度和负责任的工作态度让我在学习和生活上都受益良多。在学术上认真负责，对研究中遇到的困难和不足总是能给出指导，吴老师对学生认真负责的态度、敏锐的学术洞察力和勇于开拓的精神是我永远学习的榜样。三年里，不论是从幻灯片的展示到论文的研读，还是从生活的琐事到做人的道理，老师都会悉心指导。不论以后何去何从，吴老师都将是我一直的导师，会带着老师对我的教导和老师教会我的知识更加努力地学习生活。感谢您，老师!

其次，感谢课题组的张新鹏、冯国瑞、任艳丽和陈俊丽等老师为我们提供了良好的学习平台。感谢易标师兄在研究方向上给予我的帮助，在我有问题时，总是认真倾听耐心解答，让我收获颇多。感谢 904 实验室的柳琦云、杨天予、唐雄、陈诗怡和魏诗语等同学和 902 实验室的牛祥华、许智磊和陈驰，在实验室共同学习和研究的日子中既有温馨安静的学习环境也不乏开心快乐的交流。还要感谢我亲爱的室友武可、马晓雨、查字民，谢谢你们在这三年里对我的照顾，是你们让我的寝室生活更加丰富多彩，充满了快乐和幸福。特别感谢我的大学室友陆晨燕同学给予我精神上的支持和情感上的慰藉，一直默默地陪伴我。

此外我想感谢我的家人。感谢你们对我的辛勤养育，对我学习的支持，你们支持和鼓励一直是我继续学习的动力，因为你们我感到很幸福。祝愿您们永远平安健康幸福。

最后感谢百忙之中对我的论文进行评阅的各位专家老师们，谢谢您们对我的论文提出的宝贵建议。