利用范本构建语法模板生成仿自然语言隐写文本

苏胜君[①②]　李维斌[③]　陈　超[②]　王朔中[②]
[①](上海理工大学光学与电子信息工程学院　上海　200093)
[②](上海大学通信与信息工程学院　上海　200072)
[③](台湾逢甲大学资讯工程与电脑科学系　台中　40724)

摘　要：基于模仿函数的Mimicry文本隐写法利用上下文无关文法(CFG)构造树形结构，生成的含密仿自然文本在语义连贯性方面存在缺陷。该文提出一种改进的Mimicry文本隐写方法，从选择范本和设计语法模板两方面提高语义的连贯性和一致性。经过范本训练，在模板设计中运用置换文本库，考虑文本中各词汇、短语和句子的可置换性。模板还运用Huffman编码，充分考虑语句和词组的出现频度，以改善含密文本的自然度。这样生成的含密文本语义连贯性好，符合自然语言在字符、词汇、句法等方面的统计特性，有利于对抗人工和机器检测。

关键词：文本隐写；模仿函数；上下文无关文法；语义连贯性
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2008)08-1936-04

CFG-Based Stego-Text Generation Using Template Text and Grammar File

Su Sheng-jun[①②]　Lee Wei-Bin[③]　Chen Chao[②]　Wang Shuo-zhong[②]
[①](*College of Optical and Electronic Information Engineering*, *University of Shanghai for Science and Technology*, *Shanghai* 200093, *China*)
[②](*School of Communication and Information Engineering*, *Shanghai University*, *Shanghai* 200072, *China*)
[③](*Department of Information Engineering and Computer Science*, *Feng Chia University*, *Taizhong*, *China*)

Abstract: A previous text steganography scheme uses the mimic function and the Context-Free Grammar (CFG) to generate pseudo-natural language text to convey secret information. However, the generated text usually lacks semantic consistency and is therefore vulnerable to human evaluation. An improved mimicry technique is proposed that constructs the CFG-based grammar file from a template text that meets certain criteria so that the generated text is semantically consistent. The grammar file is constructed from a replacement text base composed of many optional words, phrases or sentences. Huffman coding is used to take into account the occurrence frequencies of various words, expressions, and syntactic structures to improve naturalness of the generated text. Thus, in addition to machine examination, the obtained stego-text is able to pass stringent human evaluation.

Key words: Text steganography; Mimic function; Context Free Grammar (CFG); Semantic consistency

1　引言

在互联网快速发展的背景下，学术界对图像、数字音频和视频中的信息隐藏进行了广泛研究，许多技术在版权保护、隐蔽通信、内容认证等信息安全领域得到应用。文本是使用最广的信息载体，形式多样，如各种文本文件、网页、电子邮件等，研究用文本作载体的信息隐藏对信息安全具有重要意义。

早在10年前Bender[1]就讨论了几种文本信息隐藏技术，包括利用单词之间空格的方法，如空1格表示0，空2格表示1；基于句法的方法如用a, b, and c表示0，用a, b and c表示1；以及利用语义的方法，例如用big表示0，large表示1。其中第1种是基于格式的，后两种则属于自然语言信息隐藏。基于格式的文本信息隐藏后来又出现了多种方法，此类技术主要是在文本域嵌入信息，而在图像域提取，其本质还是图像中的信息隐藏[2, 3]。要保证隐蔽性，修改必须很小，另一方面为了保证可靠检测，修改又不能太小，这些方法的不足之处是无法抵御重新排版或OCR攻击。为此又提出了基于字符属性或显示特征的信息嵌入[4]。

近年来人们更多地转向基于自然语言的方法，信息的嵌入和提取均在文本域利用句法和语义实现。可将隐蔽信息嵌入已有文本，也可以生成符合自然语言特征的含隐蔽信息的文本。前者可作为水印用于保护版权，也可用作隐蔽通信的一种手段，即隐写，而后者只能用于隐写。为了通过机器检测，含有隐蔽信息的文本必须符合语言学上字符、词汇和句法的多层次统计特性；另一方面，为了通过人工(阅读)检测，含密文本在语义上还必须具有连贯性和一致性。Atallah一改以往用同义词替换法直接修改载体文本嵌入水印的做法，先

2007-01-09 收到，2007-11-23 改回
上海市重点学科建设项目(T0102)资助课题

进行句法变换和语义变换，通过分析，生成具有树状结构的中间体，在其中嵌入信息后再变换回到文本域，提高了安全性[5]。Topkara等人认为自然语言水印可借用图像水印中的某些概念，例如文本的深层特性可类比于图像DCT域[6]。

另一种技术是用机器翻译生成的译文作为载体进行隐写。由于机器翻译产生的噪声和隐写引起的噪声难以区分，故可利用译文中的冗余进行隐写[7]。后来研究者又改进了译文信息隐藏协议，在提取信息时不再需要原文[8]。这一改进很重要，因为传递原文不仅占用资源，而且不安全。

本文研究仿自然语言生成的文本隐写技术，提出一种改进的Mimicry隐写文本生成方法，该方法建立在Wayner提出的模仿函数(mimic function)[9, 10]基础上，目的是改善生成文本的连贯性和自然度，从而提高隐写的安全性。

2　改进的Mimicry文本隐写方法

2.1 Mimicry算法及存在的问题

Wayner的模仿算法(Mimicry)的成功之处主要在于使隐写信息和生成的文本融为一体，根据隐写信息的内容自动产生隐写文本。这种方法在统计特性和语法两方面均符合自然语言的要求，因此能对抗隐写分析。

Mimicry隐写法采用上下文无关文法(Context Free Grammar, CFG)描述句子的结构来构造隐写文本，而通过句子结构的不同选择来实现信息隐藏。CFG中包含变量和常量，可用树形结构来表示。CFG树中有非终端和终端两种结点，可将待编码的信息存放于树的终端结点。图1是一个CFG树的实例，带下划线的粗斜体词是变量，有下一级左右两个分支。通过选择相应的终端结点代表要传递的数据0或1。CFG也可以有语法的嵌套，如**noun**的下一层也可以是从句的**subject**和**predicate**。同时还要加入符合语言习惯的限制，比如不允许无限嵌套或无限循环。

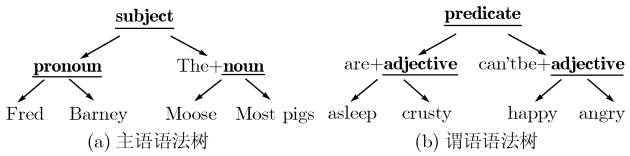

图1 用于信息隐写的上下文无关文法实例

根据图1的语法规则首先考虑主语语法树，然后再考虑谓语语法树。令图中每个分支左侧表示0，右侧表示1。若要传递秘密数据1100，则生成的隐写语句为：Most pigs are asleep。如此可产生多个句子从而形成一篇仿自然语言的文字。

Mimicry方法由基于句子的语法模板构成，但是用上下文无关文法所生成的文本无法保证句子之间语义上的连贯性，这是这种方法的不足。仅依靠语法树中的主、谓语等语法成分进行直接替换会导致上下文的主语不一致，语意混乱等问题。以下是仅按照上下无关文法生成的一段Mimicry含密文本[11]：

Dear Decision maker, Thank-you for your interest in our briefing ! We will comply with all removal requests ! This mail is being sent in compliance with Senate bill 2716, Title 9, Section 306 ! Do NOT confuse us with Internet scam artists ! Why work for somebody else when you can become rich within 83 months ! Have you ever noticed most everyone has a cellphone and nearly every commercial on television has a .com on in it ! Well, now is your chance to capitalize on this ! WE will help YOU sell more & process your orders within seconds. You can begin at absolutely no cost to you ! But don't believe us ! Mr Simpson who resides in Maryland tried us and says "Now I'm rich many more things are possible" ! This offer is 100% legal ! We BESEECH you - act now ! Sign up a friend and you'll get a discount of 50%. Thanks.

可以看出，携带秘密信息的文本虽然符合语法要求，但语义含糊，连贯性差，令人费解。用这种方法产生的文本进行隐蔽通信很容易被识别，不够安全。以下我们通过一种自下而上的设计方法，选择适当的文本作为基础构建隐写算法树，对上述方法进行改进。

2.2 范本选择

一篇成功的Mimicry隐写文本既要在字符和词汇长度两方面满足自然语言的统计特性，还应满足句法所要求的统计特性和语义上的前后连贯性。如此生成的文本虽然在内容上似乎没有实际意义，却可以混在大量存在的垃圾信息中传播而不易引起注意[12]。为了更好地符合语言学上语义的连贯性和基于句法的统计特性，我们设计一个含有一定上下文相关信息的树状语法文件，用于生成隐写文本。为此首先选定语法树底层的文本样板，将它称为范本，然后以此为基础由下而上，构建具有良好语义连贯性的隐写算法树。

为了加强语义连贯性和上下文相关信息，范本要遵从以下几项原则：

(1)选定的文本在词汇、句子、语义3个层次上都必须符合一定的语言规则和统计特性。

(2)在文本内容方面，以文学性、描述性或介绍性文字为主，避免观点尖锐的评论文字。这样的文本往往具有较强的上下语义包容性。

(3)尽量包含数字和可替换的列举项目。这是因为数字或列举项目的改变不易对文本的语意和连贯性造成明显的影响，便于增加信息嵌入率。例如用一场足球比赛的现场评论作为隐写模板时，将“3号球员射门”改为“5号球员射门”，既可实现信息隐藏，又不会引起上下文脱节，确保不影响语意连贯性。

2.3 隐写算法树的构建

为了避免语法树中上下文无关的弊端，必须充分考虑范本的特点。通过对范本进行训练学习，最终生成置换文本库和语法模板。

(1)置换文本库　首先要考虑文本中各词汇、短语和句子的可置换性，生成一个置换文本库。它类似于同义词库，但跟同义词库相比，置换文本库的含义更广泛，更灵活。例如{a comprehensive university, a beautiful school, a famous university, …}可作为置换短语库的子集；{rich diversity of majors, many disciplines, a good number of majors, …}构成另一子集。

(2)同位语用于隐藏信息　当文本中出现并列结构的词汇、短语、句子或段落时，可将他们按任意顺序排列用于嵌入信息。假设有n个并列短语，则一共有p_n^n种不同排列，可用于嵌入$\log_2(p_n^n-1)$比特信息。实际应用时可考虑语义和统计特性等因素对某些选择加以限制。

(3)语法模板　根据范本的语法结构和置换文本库，生成一个语法模板文件。在生成模板文件时采用Huffman树状结构，编码长度由相应词汇、短语和语句的出现频度来决定。在自然语言信息隐藏中应用Huffman编码可使含密文本比较容易通过机器测试，因而较为安全[13]。考虑到二进制编码的需要，将每个结点下的分支数设为2的幂次。

这种隐写算法能在以下几个方面保证句子之间的语义连贯性和自然度。首先是选取典型的自然语言文本作为范本，它符合语法和语义多个层次的要求。另外范本经过自下而上地训练学习，由此得到隐写算法。这样生成的密写文本能保持原范本中句子之间连贯性。其中同位语结构和置换文本不会改变语法和语义内容，语法模版也不会损失语义的连贯性。为了克服原Mimicry算法树形成后难以扩充的缺点，在改进算法中设计的置换文本库、同位语和语法模版是相对独立的，它们可以在训练的基础上，不断得到扩充和改善，进一步加强句子的连贯性和语义的自然度。

2.4 密写和信息提取方法

置换文本库和语法模板文件可通过适当的渠道传送给接收方，为保证安全性应使用密钥。发送方根据待传递的秘密信息(二进制序列)、模板文件和置换文本库生成隐写文本。接受方收到到隐写文本后，把它同模板文件、置换文本库对照即可提取秘密信息。

为了增强安全性，在隐写前可以先对数据进行加密。隐写长度小于文本的可嵌入量时可随机确定隐写位置，从文本的中部开始嵌入信息，并把隐写位置作为密钥。为了避免编码二义性，也可适当加入冗余码进行检错和纠错。

3 实验及结果分析

3.1 mimicry密写实例

我们分别选取一段某大学的介绍性文字和文学作品段落为范本生成模板。文本中的词汇、短语、句子都可以按照置换文本库进行替换，以生成不同的隐写信息。下面是摘选的一段文学类范本：

But you can see, the phrase “love and death” can be produced in two ways. This is a big headache for the parser and you should avoid this situation at all costs. Right now the parser will notice when it encounters ambiguities and signal an error, but in some sense this is too late. I am considering building in a function that will determine whether a certain grammar can lead to ambiguity, but I haven’t done it yet. It will probably not be general because it is undecidable whether a particular language defined by a grammar is always ambiguous in all manifestations …

是用自下而上的方法，从上述范本出发设计的语法模板文件见图2。其中树结构中的结点分为终端结点和非终端结点。带下划线的粗体字符表示非终端结点，下层仍有分支；斜体字符表示终端结点，为树的分支末结点，不再有分支。

Start → *but* **things**, *the phrase* **phrase** *can be produced in* **number ways**. **Next** *you should avoid this situation at all costs.* …
things → **name helpverb motion**
next → *This is a big headache for the parser* **conjective name**
name → *you* ||*he* ||*they*|| *we*
helpverb → *can* ||*may* || *could* || *should*
motion → *see* ||*sight* || *look* || *glom*
phrase → “*love and death*” || “*life and hate*” || “*love and hate*” || “*love and life*”
numberways → *a few* **ways**||*many* **ways** ||*two* **ways**|| *several* **ways**
ways→*ways* || *factors* || *thoughts* || *directions*
conjective → *and* || *while*
…

图2 采用自下而上的方法设计的语法模板树文件

如：要密写的数据为0101011000001000…，则根据语法模板文件可自动生成文本如下：

But he may sight, the phrase “love and hate” can be produced in a few ways. This is a big headache for the parser while you should avoid this situation at all costs …

3.2 并列结构和Huffman编码树实例

图3是一段介绍类文本的语法模板，其中运用了数字、Huffman编码和并列结构等技术。

采用Huffman编码的树形置换文本库见图4。在每个文法中终端结点(频度较大或语意较佳)位于分支的左端，即嵌入

Zhonghua University is a **adjective1** *university with* **phrase1 noun3**, *it consists of* **number1 noun1**, …
These colleges and departments are:
College of Sciences
College of Fine Arts
College of International Business and Management
School of Communication and Information Engineering
…

图3 采用自下而上的方法设计的语法模板文件

的信息0；非终端结点嵌入信息1，并进入下层文法。例如要隐藏比特流1011110110 …，则生成的语句是：

Zhonghua University is a beautiful university with many disciplines. It consists of 24 colleges …

These colleges and departments are:

School of Communication and Information Engineering

College of Fine Arts

College of Sciences

College of International Business and Management

…

模板中20多个学院是一种并列结构，可嵌入的数据量为 $\log_2(p_{20}^{20}-1)\approx 61$ bit。

adjective1→*comprehensive* || **adjective2**
phrase1→*rich diversity of* || **phrase2**
phrase2→*a good number of* || *many*
noun3→*majors* || *disciplines*
noun1→*colleges and schools*||**noun2**
noun2→*colleges*||*schools*
adjective2 →*beautyful*||*great*
number1 →*20*||**number2**
number2 →*22*||*24*||*21*||*23*

图4 置换文本库和Huffman编码的树形结构

对比本节生成的文本和第2节给出的例子，可见这里所提出的文本密写算法改善了语义的连贯性，所产生的文本更加符合自然语言习惯，从而有利于对抗人工和机器检测，提高了密写的安全性。

4 结束语

随着信息安全日益受到重视，信息隐藏技术发展很快。与图像、语音等媒体的数据隐藏相比，自然语言文本中数据隐藏技术的发展相对滞后。常用的自然语言文本隐写算法如Mimicry方法采用上下文无关文法，使得隐写文本的语义连贯性不强，成为一个明显的安全漏洞。本文对此进行了改进，从底层开始选定语义连贯性强的范本，通过分析范本的词汇和句法结构生成隐写模板，用于自动生成含秘密信息的隐写文本。所得到的含密文本符合英语自然语言的字符、词汇、句法的多层次统计特性，对它进行隐写分析难度较高，因此具有良好的安全性。这种方法还可实现较大地信息嵌入量。

该方法的性能可通过考虑同位语的不同频度、避免语法模板文件可能产生的二义性、应用语法嵌套等方面实现进一步的改善。还可考虑在Mimicry方法中引入上下文相关算法，实现更符合语言学要求的文本隐写算法。

参 考 文 献

[1] Bender W, *et al.*. Techniques for data hiding. *IBM Systems Journal*, 1996, 35(3,4): 313–336.

[2] Brassil J T, Low S, and Maxemchuk N F. Copyright protection for the electronic distribution of text documents. *Proce*, *IEEE*, 1999, 87(7): 1181–1196.

[3] Takizawa O, *et al.*. Method of hiding information in agglutinative language documents using adjustment to new line positions. LNCS/LNAI, 2005, 3683: 1039–1048.

[4] 肖湘蓉，孙星明. 基于内容的英文文本数字水印算法设计与实现. 计算机工程，2005, 31(22): 29–31.
Xiao Xiang-rong and Sun Xing-ming. Design and implementation of content-based english text watermarking algorithm. *Computer Engineering*, 2005, 31(22): 29–31.

[5] Atallah M, *et al.*. Natural language watermarking and tamperproofing. fifth information hiding workshop, IHW'02, Noordwijkerhout, The Netherlands, LNCS, 2002, 2578: 7–9.

[6] Topkara M, *et al.*. Natural language watermarking. Proceedings of the SPIE, International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI, San Jose, CA, USA, January, 2005: 17–21.

[7] Grothoff C, *et al.*. Translation-based steganography. Proceedings of Information Hiding Workshop (IH2005), Barcelona, Spain, 2005: 213–233.

[8] Stutsman R, *et al.*. Lost in just the translation. The 21st Annual ACM Symposium on Applied Computing, Dijon, France, April, 2006: 23–27.

[9] Wayner P. Mimic functions. *Cryptolgia*, 1992,16(3):193-214.

[10] Bennett K. Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text, CERIAS Tech Report, 2004–13.

[11] Peter Wayner. encode and decode. http://www. spammimic. com.

[12] Gaudin S. Record Broken: 82% of U.S. Email is Spam. May 5, 2004. http://itmanagement.earthweb.com/secu/article.php/3349921.

[13] Bergmair R. Natural language steganography and an AI-complete security primitive. Proceedings of the 7th Information Security Conference, LNCS, Springer Verlag, September 2004, 3225: 257–267.

苏胜君：　女，1970年生，博士生，讲师，研究方向为信息与通信工程.

李维斌：　男，1970年生，博士，副教授，研究方向为信号与信息处理.

陈　超：　男，1965年生，博士生，高级工程师，研究方向为信息与通信工程.

王朔中：　男，1943年生，博士，教授，博士生导师，研究方向为信息与通信工程.