基于语义的文本隐藏方法①

Text Steganography Based on Semantics

徐迎晖　杨 榆　钮心忻　杨义先　（北京邮电大学信息安全中心 100876）

摘要：基于语义的文本隐藏是一种通过变换字词本身而不改变语义的隐藏方法，目前主要采用同义词替换来实现。文章提出一种汉字同音替换的基于语义的文本隐藏方法，在隐蔽性和抗攻击性上有一定的改进，并可通过对句子进行分词和词性标注的方法实现机器自动隐藏和盲检测。

关键词：语义　文本隐藏　信息隐藏　同音替换

1　引言

随着计算机和通信网技术的发展与普及，数字音像制品以及其他电子出版物的传播和交易变得越来越便捷，人们如今可以通过因特网发布自己的作品、私有信息等，但是随之而出现的问题也十分严重：如作品侵权更加容易，篡改也更加方便。如何既充分利用因特网的便利，又能有效地保护知识产权，已受到人们的高度重视。同时网络信息的高速增长也为秘密信息提供了大量载体，使得降低秘密信息被截获的概率更为有效。

2　文本隐藏方法概述

目前比较常用的隐藏介质是图像或声音，它们包含了大量的冗余信息，由于人的视觉和听觉特性中存在一些不敏感点，我们可以利用这些特点在介质里隐藏信息而不被察觉。与图像和声音相比，在文本里面隐藏信息是比较困难的，它只包含非常少的冗余信息，目前的文本隐藏方法可分为以下三类。

（1）基于格式的文本隐藏方法。这种方法针对的是具有一定排版格式或文件结构的文本，相对于纯文本而言，格式化文本的格式信息中包含了较多的冗余量，人眼无法辨别出排版格式的一些细微变化，普通用户通常也不会检查文件结构中发生的变化。现有的基于格式的文本隐藏方法主要有调整行间距、字间距、字体、字符大小，构建字符特征编码，修改文件结构中的标记、属性或预留字段等。这些方法只考虑保留文本的视觉形式而不考虑其具体内容，通用性较好，隐藏容量也较大，但是抵抗重新排版或查看源代码的能力较弱。

（2）基于句法的文本隐藏方法。句法研究词如何排列组成正确的句子，每个单词在句子中的结构角色及短语之间的构成关系。基于句法的文本隐藏方法改变措辞和句子结构而不显著改变句子意思和语气。这类方法包括修改含混标点、句子分拆和组合、移动附加语位置、加入删除形式主语、主动被动语态变换、加入删除冗余短语等[2]。这类方法的隐蔽性较好，但是受到文本写作风格和内容的影响，在达到较好自然度的同时隐藏容量受到限制。

（3）基于语义的文本隐藏方法。语义研究词语的意义以及在句子中词语意义是如何互相结合以形成句子意义。基于语义的文本隐藏方法主要改变的是词语本身，以同义词替换的方法为代表。同义词替换的隐蔽性较好，但是同义词之间存在的细微差别体现在上下文中时有时会比较明显，频繁的替换也会影响到文本写作风格的一致性，因此要达到较好自然度隐藏容量会受到限制。

① 本课题得到国家重点基础研究发展规划项目（TG1999035804）、北京市自然科学基金项目、教育部优秀青年教师资助计划项目资助

在提取隐蔽信息时需要提供嵌入时使用的同义词表或对照原文,较难实现盲检测。

3 基于语义的文本隐藏方法

汉语中除了大量的同义词之外,还存在大量的同音替换现象,如假借、通假、异形词。这些词与同义词的区别在于,它们在发音上、意义上完全相同而只是书写形式不同,而且具有固定的替代形式。因此没有同义词之间意义上的细微差别和同义词组类的不一致性。利用假借字、通假字、异形词之间的同音替换可以进行信息隐藏。

3.1 隐藏原理

(1) 汉语发展过程中出现了大量的假借字、通假字、异形词和异体字,其中很大一部分在现代汉语中已很少使用或被规范整理淘汰。文献[6]列举了 338 组常用的异形词及其推荐词形,这些词对可用于同音替换隐藏。例如利用词对"贤惠——贤慧"进行同音替换:

原句为:看雪芹在《红楼梦》里总提起宝钗很贤惠,似乎怕人一不小心就会忘记,把她当作奸人。

替换为:看雪芹在《红楼梦》里总提起宝钗很贤慧,似乎怕人一不小心就会忘记,把她当作奸人。

这种隐藏的处理方法与同义词替换法相似,但有两点区别。一是同音替换词表中的各项组成是基本固定的,而同义词表的各项组成对于不同用户会有较大差异。二是对于同义词词对只在某个词性下成立的情况,还需要对同义词进行词性判定。运用这两类方法进行替换时,考虑到汉语句子的词之间没有分割标志,待替换的词与相邻字的组合上可能存在词语组合歧义,若要实现机器自动隐藏和检测,则需先对待处理文本进行分词和切分歧义消除。

(2) 上面提到的同音替换词对在文本中出现的频率有一定的限制,另一种较为通用的替换方式是采用结构助词词对"的——地"和"的——得"。

"地"作结构助词时用在状语后,表示状语和中心词之间的修饰关系,与"的"可作同音替换[7]。其中一些典型结构形式为:(副词、形容词) + 地 + (动词、形容词)。

"得"作结构助词用在动词或形容词后面,连接表示程度或结果的补语,或用在动词和补语中间表示可能时,与"的"可作同音替换[7]。其中一些典型结构形式为:(动词、形容词) + 得 + (动词、副词、形容词)。

通过选择"的——地"和"的——得"词对中的前者或后者可分别嵌入1、0 信息。

利用词对"的——地"进行替换的例子:

原句为:预计公司今后几年专利申请的数量将会更加快速地增长。

替换为:预计公司今后几年专利申请的数量将会更加快速的增长。

利用词对"的——得"进行替换的例子:

原句为:当宇航员打开舱门的一刻,等待已久的人们都高兴得跳了起来。

替换为:当宇航员打开舱门的一刻,等待已久的人们都高兴的跳了起来。

采用结构助词词对"的——地"和"的——得"进行隐藏时,需要判定"的地得"在句中的词性及相临中心词的词性,只有符合替换条件的才能进行。每个符合替换条件的结构可嵌入1比特信息。两例句中第一个"的"字位于定语(名词和名词性短语)和中心词(名词)之间,不符合替换条件,故不能用于隐藏信息,应保持为原文状态。词性判定可以人工完成,也可以通过自然语言处理中的句子分词和词性标注算法来实现,利用后者可以实现机器自动隐藏和盲检测。

3.2 嵌入方法

前一种同音替换法的密文信息嵌入方法和同义词替换法相同,不再赘述。后一种结构助词词对同音替换法的密文信息自动嵌入可以按照以下步骤进行。

(1) 将密文转换成二进制码序列。

(2) 搜索包含"的"、"地"或"得"的句子。

(3) 使用基于规则或基于统计的方法对该句进行自动分词。若分词后"的"、"地"或"得"与其它字一起组成词,或"的"、"地"或"得"处于句首句尾,则这些情况不符合替换条件,只需处理在分词后以单字形式出现在句中的"的"、"地"或"得"。若不存在这种情况,则回到(2)继续搜索。

(4) 对单字形式出现在句中的"的"、"地"或"得"及它们前后相邻的中心词进行词性标注,词性标注有基于规则、统计、机器学习、神经网络或混和的方法。

这一步并不需要完成句中所有词的词性标注工作。

(5) 若词性标注结果存在符合结构"(副词、形容词)+地或的+(动词、形容词)"或"(动词、形容词)+得或的+(动词、副词、形容词)"的情况,则该处可进行嵌入。否则回到(2)继续搜索。

(6) 先对原文进行规范化处理,将"(副词、形容词)+地或的+(动词、形容词)"、"(动词、形容词)+得或的+(动词、副词、形容词)"分别规范为"(副词、形容词)+地+(动词、形容词)"、"(动词、形容词)+得+(动词、副词、形容词)",对比规范化文本嵌入密文信息。当前待嵌入密文的二进制位为 0 时,保持"地"或"得"不变;为 1 时将"地"或"得"替换成"的"。

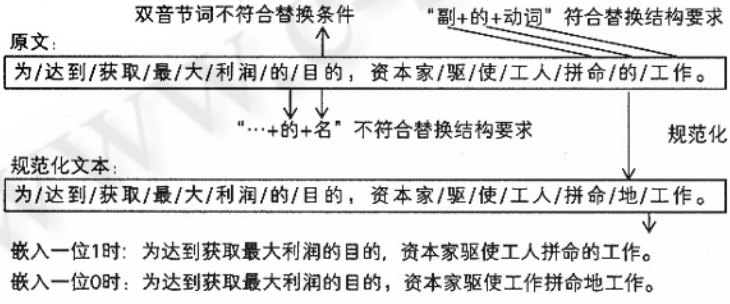

图 1　密文信息嵌入示例

(7) 转步骤(2)继续嵌入密文的下一个二进制位,直至密文已嵌入完毕或原文已搜索完毕。

我们以句子"为达到获取最大利润的目的,资本家驱使工人拼命的工作。"为例,说明一下上述密文信息嵌入过程。

该句的前后两段均包含"的"、"地"或"得",故先对它们进行句子分词和词性标注,按照北大版汉语词性标记集的词法分析结果为:

为/p 达到/v 获取/v 最/d 大/a 利润/n 的/u 目的/n ,/w 资本家/n 驱/vg 使/v 工人/n 拼命/d 的/u 工作/v 。/w 其中第一个和第三个"的"字均以单字形式出现在句中,有待进一步确定是否符合替换条件;第二个"的"字与其它字一起组成词,且出现在句尾,故不符合替换条件。

下面来确定第一个和第三个"的"字是否符合替换条件。前者与其相邻中心词的结构为"n + u + n",即"名词+助词的+名词",不满足替换条件;后者与其相邻中心词的结构为"d + u + v",即"副词+助词的+动词",符合替换条件。

这样,整句存在一个符合替换条件的结构,故该句可嵌入 1 比特信息。我们先对原文中符合替换条件的结构进行规范化处理,再根据待嵌入密文比特是 1 还是 0 分别对规范化文本进行"地(得)—>的"替换或是保持不变,最后得到嵌入了 1 比特信息的文本。如图 1 所示。

3.3 提取方法

后一种结构助词词对同音替换法的密文信息自动提取可按以下步骤进行,除对比规范化文本的步骤不同外,它和嵌入方法基本类似。其中所述的原文指已嵌入密文的文本。

(1) 搜索包含"的"、"地"或"得"的句子。

(2) 使用基于规则或基于统计的方法对该句进行自动分词。若分词后"的"、"地"或"得"与其它字一起组成词,或"的"、"地"或"得"处于句首句尾,则这些情况不符合替换条件,只需处理在分词后以单字形式出现在句中的"的"、"地"或"得"。若不存在这种情况,则回到(1)继续搜索。

(3) 对单字形式出现在句中的"的"、"地"或"得"及它们前后相邻的中心词进行词性标注,这一步并不需要完成句中所有词的词性标注工作。

(4) 若词性标注结果存在符合结构"(副词、形容词)+地或的+(动词、形容词)"或"(动词、形容词)+得或的+(动词、副词、形容词)"的情况,则该处可进行提取。否则回到(1)继续搜索。

(5) 先对原文进行规范化处理,将"(副词、形容词)+地或的+(动词、形容词)"、"(动词、形容词)+得或的+(动词、副词、形容词)"分别规范为"(副词、形容词)+地+(动词、形容词)"、"(动词、形容词)+得+(动词、副词、形容词)",将原文与规范化文本进行对比,若"地"或"得"保持不变,则原文嵌入了一个二进制位 0;若"的" 被规范化为"地"或"得",则原文

嵌入了一个二进制位1。

(6) 转步骤(1)继续提取下一个二进制位,直至原文已搜索完毕。

3.4 分析

(1) 隐藏容量。本隐藏方法的容量受到语料特性的影响,很难得到一个统一的准确值,只能根据统计结果作出一个大致的估计和参考。文献[5]列出了结构助词"的(·de)"、"地(·de)"、"得(·de)"的一些统计数据,如表1所示,统计语料未包括诗歌韵文、古代汉语和外国作品的翻译文章。由表中数据可知,"地(·de)"和"得(·de)"的总计出现频率为0.97767%,这个出现频率可以用来估计结构助词词对"的——地"和"的——得"同音替换方法的隐藏容量,即约102个汉字可隐藏1比特的信息。这个容量适合应用于文本水印等较小数据量的领域。

表1 "的地得"词频统计

汉字	出现频率	词次等级	使用度等级
的(·de)	5.38720%	1	1
地(·de)	0.63043%	17	17
得(·de)	0.34724%	34	34

(2) 句子分词和词性标注。本隐藏方案中,句子分词和词性标注算法是实现机器自动嵌入和提取的关键。汉语自动分词和词性标注包括基于规则和基于统计的方法,通常为提高正确率,采用基于规则和基于统计方法的混和策略。分词和词性标注的过程也互相结合,以消除切分歧义和词性兼类歧义。一些研究结果表明[4],采用分词和词性标注一体化方法得到的正确率可在94%左右。

考虑到本隐藏方案并不需要对全句进行分词和词性标注,只需检测是否存在符合替换要求的特定结构,因此对分词和词性标注算法的难度要求有所降低,现有的正确率有望进一步提高。

考虑在嵌入和提取时,若采用同样的分词和词性标注算法,那么发生误判的情况是基本一致的,这种情况对于嵌入和提取信息也是有利的。

(3) 抗攻击能力。结构助词词对"的——地"和"的——得"同音替换文本隐藏方法对于句内修改、增删攻击有较强的抵抗力。这两种替换符合语法和人们的日常习惯,不会引起怀疑,其隐蔽性很好。在这两种替换中,结构助词和相邻词的结合非常紧密,删除或增加该结构助词后将不符合语法。同一句中也极少出现两个以上符合替换要求结构。

对于抵抗增删少量句子攻击的能力也较强,主要体现在其隐藏容量小和分布较均匀上,针对增删少量句子的攻击,可以在嵌入前对密文先施加一定的冗余编码。对于增删大篇幅段落的的攻击,运用文献[3]纠错交织分帧技术将获得一定的抵抗能力。

4 结语

本文提出一种基于语义的汉字同音替换文本隐藏方法,着重研究了结构助词词对"的——地"和"的——得"的同音替换方法。同音替换方法的隐蔽性和抗攻击性较经典的同义词替换方法有较大的改进,并可通过自然语言处理中句子分词和词性标注的方法实现机器自动隐藏和盲检测。但它的隐藏容量较小,适合于文本水印等较小隐藏数据量的领域。典型的应用载体包括非诗歌韵文的中长篇著作、短篇合集等。这些载体的发行量大、流传广泛,对运用本方法具有较好的实用价值。

参考文献

1 W. Bender, D. Gruhl, N. Morimoto. Techniques for data hiding [J]. IBM Systems Journal, 1996, 35 (3&4):313 -336.

2 张宇、刘挺、陈毅恒等,自然语言文本水印[J],中文信息学报,2005,19(1):56 -62,70。

3 白剑、杨榆、徐迎晖等,基于文本的信息隐藏算法[J],计算机系统应用,2005,4:32 -35。

4 赵铁军等,机器翻译原理[M],哈尔滨工业大学出版社,2000.6。

5 北京语言学院语言教学研究所,汉语词汇的统计与分析[M],外语教学与研究出版社,1985.4。

6 第一批异形词整理表,中华人民共和国教育部,国家语言文字工作委员会,2002.3.31。

7 高级汉语词典,海南出版社,2003.11。