第 45 卷 第 3 期 计 算 机 工 程 2019 年 3 月
Vol. 45 No. 3 Computer Engineering March 2019

·开发研究与工程应用· 文章编号：1000-3428(2019)03-0309-06 文献标志码：A 中图分类号：TP391

基于 Word2vec 的自然语言隐写分析方法

喻靖民[a,b]，向凌云[a,b,c]，曾道建[a,b]

（长沙理工大学 a. 综合交通运输大数据智能处理湖南省重点实验室；b. 计算机与通信工程学院；
c. 智能道路与车路协同湖南省重点实验室，长沙 410114）

摘 要：为数字化表示文本内容的语义信息，并提高基于同义词替换的隐写文本检测精度，提出一种新的自然语言隐写分析方法。利用 Word2vec 对大规模语料库进行训练获得包含丰富语义信息的多维词向量，使用同义词及其上下文词向量之间的余弦距离度量 2 个词之间的相关度，并计算同义词在特定上下文中的合适度。根据信息嵌入过程中同义词替换操作对文本同义词合适度的影响提取检测特征形成特征向量，采用贝叶斯分类模型训练特征向量得到隐写分析特征，从而识别隐写文本。实验结果表明，该方法对于不同嵌入率下隐写文本的平均检测精确率和召回率分别达到 97.71% 和 92.64%，具有较好的检测性能。

关键词：自然语言；词向量；同义词替换；隐写分析；上下文合适度

中文引用格式：喻靖民，向凌云，曾道建. 基于 Word2vec 的自然语言隐写分析方法[J]. 计算机工程，2019，45(3)：309-314.
英文引用格式：YU Jingmin，XIANG Lingyun，ZENG Daojian. Natural language steganalysis method based on Word2vec[J]. Computer Engineering，2019，45(3)：309-314.

Natural Language Steganalysis Method Based on Word2vec

YU Jingmin[a,b]，XIANG Lingyun[a,b,c]，ZENG Daojian[a,b]

（a. Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation；b. School of Computer and Communication Engineering；c. Hunan Provincial Key Laboratory of Smart Roadway and Cooperative Vehicle-Infrastructure Systems，Changsha University of Science and Technology，Changsha 410114，China）

【Abstract】In order to represent the semantic information of the text content for digitization and improve the accuracy of detecting stego texts based on synonym substitution，a novel natural language steganalysis method is proposed. Word2vec is employed to train a large-scale corpus to obtain multi-dimensional word vectors which contains rich semantic information. Then，it uses the cosine distance between a synonym and its context word vector to measure the correlation between two words，and calculates the fitness of synonyms in a specific context. According to the effect on the context fitness of the synonyms caused by the synonym substitutions in the embedding process，detection features are extracted to form a feature vector，and the Bayesian classification model is employed to train feature vector for the task of steganalysis feature to detect the stego texts. Experimental results show that the proposed method has good detection performance，whose average detection precision and average recall for the stego texts with different embedding rates achieve 97.71% and 92.64%，respectively.

【Key words】natural language；word vector；synonym substitution；steganalysis；context fitness
DOI：10.19678/j.issn.1000-3428.0050407

0 概述

近年来，研究人员提出了各种自然语言隐写方法[1-2]。这些方法通过语言等价变换将秘密信息嵌入到自然文本中达到隐蔽通信的目的，然而不法分子有可能利用这些方法从事非法活动。为有效监管和阻止自然语言隐写方法的滥用，出现了一种自然语言隐写分析技术，检测文本中是否存在秘密信息。最早使用的自然语言隐写方法是利用同义词之间的替换实现秘密信息的隐蔽嵌入。文献[3]提出一个基于同义词替换的隐写系统 T-Lex，该系统简单实用，但容易导致语法错误以及统计特性的改变。为解决这些问题，国内外学者提出了多种新方法，如文献[4]通过使用 Google *n*-gram 语料库计算单词的上下文合适度动态选择用于替换的同义词，提出基于顶点编码的隐写方法。文献[5]提出基于矩阵编

基金项目：国家自然科学基金(61202439，61602059)；湖南省教育厅科学研究重点项目(16A008)。
作者简介：喻靖民(1993—)，男，硕士研究生，主研方向为隐写分析、自然语言处理；向凌云、曾道建，讲师、博士。
收稿日期：2018-02-05 **修回日期：**2018-03-05 **E-mail：**243845445@qq.com

码的同义词替换隐写方法减少载体文本的修改量。文献[6]通过二元依存关系获取最佳同义词替换集，设计新的隐写方法。

尽管使用同义词替换的隐写方法可以使替换前后文本的语意基本保持一致，但不可避免地会使隐写前后文本存在统计上的差异，从而被其他隐写分析技术利用检测出隐藏信息[7]。文献[8]利用上下文的联合频率信息评估同义词在上下文中的合适度，通过合适度相关序列的统计特征检测隐藏信息。文献[9-10]提出基于相对频率分析、基于上下文词聚类分析的检测方法，分别计算同义词词频的均值和方差，以及利用同义词与上下文单词的共现频率估算同义词合适度提取检测特征，进一步提高检测精度。文献[11]利用同义词之间词频的差异，从同义词的词频大小和同义词个数定义同义词的属性，从而提取相应的隐写分析特征。上述方法使用的检测特征主要基于词频特性，属于浅层特征，不能较好地表达语义语法和其他深层语言特征，影响了隐写分析的检测精度。

本文提出一种基于 Word2vec 的自然语言隐写分析方法，利用分布式词表示工具 Word2vec 为每个单词训练出包含丰富语义信息的多维词向量，使用同义词及其上下文词向量之间的余弦距离度量2 个词之间的相关度，以此计算同义词在特定上下文中的合适度并提取检测特征，同时将检测特征输入到贝叶斯估计模型中进行训练和测试。

1 自然语言隐写分析方法

1.1 基于 Word2vec 的词表示

在对文本进行分析前，先将文本内容转换成数字(即词向量化)，再进行后续运算。词向量化提供了一种数学化的方法，将自然语言符号信息转化为向量形式的数字信息。文献[12]提出词的分布式表示概念，将单词表示为维度较低的稠密向量且该向量能够刻画单词语义之间的相似度。词的分布式表示能克服另一种典型的词向量表示模型 One-hot[13] 的缺点，如向量维数灾难及不能较好地刻画词与词之间的相似性。

词的分布式表示的实现方法有很多。文献[13-14]提出 2 个基于神经网络的语言模型，并具体实现为词向量工具 Word2vec。Word2vec 是一个开源工具，通过对大规模语料库的训练，能够将单词表示成包含丰富语义信息的多维实数向量。

Word2vec 采用 2 种神经网络语言模型 CBOW 和 Skip-gram 实现词的分布式表示。这 2 种模型均属于浅层的双层神经网络，包括输入层、隐藏层和输出层，结构如图 1 所示。

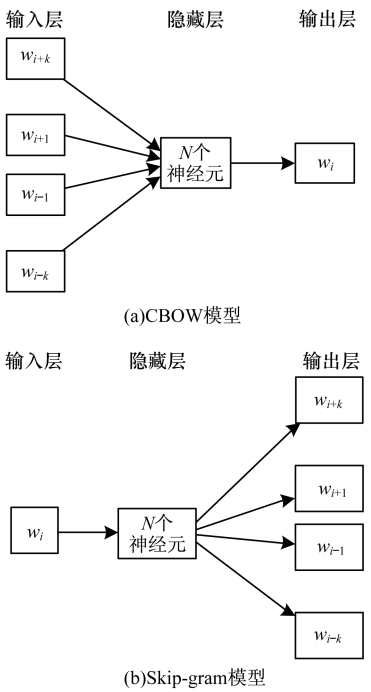

图 1 Word2vec 模型架构

CBOW 模型的建模思想是根据上下文预测当前词语的概率，即已知上下文 $w_{i-k},\cdots,w_{i-1},w_{i+1},\cdots,w_{i+k}$ 的词向量，预测当前词 w_i 的词向量。而 Skip-gram 模型刚好相反，根据当前词语来预测上下文的概率，即已知当前词 w_i 的向量预测其上下文 $w_{i-k},\cdots,w_{i-1},w_{i+1},\cdots,w_{i+k}$ 的词向量。神经网络语言模型在大规模语料库上反复运行上述过程进行训练，获得最终的词向量，Skip-gram 模型训练速度较慢，CBOW 模型训练速度相对较快，但对于低频词 Skip-gram 模型的词向量效果较好。在基于同义词替换的隐写方法中，使用的同义词多数是使用频率较低的词，针对这类隐写文本的检测，为获得更好的隐写分析检测结果，选用 Skip-gram 模型训练词向量。

Skip-gram 模型的训练目标是找到对于预测句子或文档中的上下文词语有用的词表示，本质上是一种词袋模型，基于单词的上下文训练词向量。每个词向量反映了上下文单词的加权值，通过训练学习到的向量能较好地表征一个词的语义信息。因此，本文利用 Skip-gram 模型从大量非结构化语料库文本数据中学习到高质量的词向量，将需要量化的单词映射到一个多维的向量空间。获得的词向量能够有效地揭示词之间深层和隐含的语义关系，如词与词之间的逻辑关系、同义词及其上下文词之间的相关性等。特别地，词向量能较好地刻画词语之间的语义相似度，如 vector(‘Paris’) – vector(‘France’) + vector(‘Italy’) 的结果与 vector(‘Rome’) 非常相近，而 vector(‘king’) – vector(‘man’) + vector(‘woman’) 与 vector(‘queen’) 非

常接近[13]。目前,此类词向量已广泛用于自然语言处理中的许多任务,如情感分析[15]、实体关系抽取[16]等,并取得了较好的效果。

给定一个单词 w,通过 Word2vec 工具选用 Skip-gram 模型获得的词向量表示如下:

$$\boldsymbol{V}(w)=\{v_1,v_2,\cdots,v_m\} \tag{1}$$

其中,m 表示词向量的维数,其值在训练前预先设定。一般而言向量维数越大越好,但同时向量的维数越大计算复杂度越高[13],过小则包含的语义信息较少,会降低词向量的表示能力,因此应当设置适中的 m 值。

1.2　同义词的上下文合适度计算

定义 1(可替换元素) 在基于同义词替换的隐写方法中,将用于嵌入信息的同义词定义为可替换元素。每个可替换元素存在 1 个或多个与其意思相近的可相互替换的元素。

定义 2(替换集) 可相互替换的所有可替换元素的集合,在基于同义词替换的隐写方法中一个同义词组即为一个替换集。如可替换元素 $ss_1,ss_2,\cdots,ss_x$ 是可相互替换嵌入信息的具有近似含义的所有同义词的集合,是一个替换集,其中 x 表示替换集合中元素的总数。每个可替换元素唯一地属于一个替换集合。如 {timeworn, hackneyed, trite} 表示一个具有 3 个可替换元素的替换集。

定义 3(上下文) 对于文本中指定的一个词,其前后一定窗口范围内的单词集合称为该词的上下文。假设窗口的大小为 k,当前词为 w_i,则该词前后 k 个单词为当前词的上下文,表示为 $c(w_i)=\{w_{i-k},\cdots,w_{i-1},w_{i+1},\cdots,w_{i+k}\}$。

由于词向量包含了丰富的语义信息,能够刻画 2 个词之间的相似性,因此利用词向量度量同义词与其上下文中词的相关度。

定义 4(词相关度) 2 个词对应的词向量之间的余弦距离定义为 2 个词的词相关度。

假定词 w 和 s 的词向量分别为 $\boldsymbol{V}(w)$ 和 $\boldsymbol{V}(s)$,根据定义 4,w 和 s 的词相关度为:

$$sim(s,w)=\frac{\boldsymbol{V}(s)\cdot\boldsymbol{V}(w)}{\|\boldsymbol{V}(s)\|\cdot\|\boldsymbol{V}(w)\|} \tag{2}$$

在文本中判断一个词使用是否恰当,通常情况下该词周围的词中距离越远的词对该词的影响力越小,一般只会是一定距离内的词决定了该词的使用是否合适。假设只有一定上下文窗口内的词才会影响可替换同义词在固定上下文中的合适度,因此利用上下文词与可替换同义词之间的词相关度衡量可替换同义词在当前上下文中的合适度。通过观察同义词的上下文合适度判断同义词是否被替换。

定义 5(上下文合适度) 单词 s 及其上下文中各单词的词相关性之和定义为词 s 的上下文合适度。

假定同义词 s 的上下文为 $c(s)=\{w_1,w_2,\cdots,w_k,\cdots,w_{2k}\}$,词 s 与上下文中词 w_i 的词相关度为 $sim(s,w_i)$,根据定义 5,则 s 在当前上下文的合适度为:

$$F=\sum_{i=1}^{2k}sim(s,w_i) \tag{3}$$

1.3　隐写分析特征提取

对于待分析文本中出现的所有同义词,可以根据定义 5 计算上下文合适度。一般而言,正常文本中所使用的同义词是与当前上下文最合适的词,因此将该同义词对应替换集合中的可替换元素依次放入当前上下文中,计算出相应的上下文合适度后,经过比较可发现,原始使用的同义词上下文合适度通常是其替换集中元素所对应上下文合适度的最大值。然而,经过同义词替换操作进行信息嵌入后,很大程度上替换后元素相应的上下文合适度要比替换前原始使用同义词的上下文合适度小,具体示例如下:

原始例句:A newspaper should be lively and should avoid hackneyed stuff。

隐写例句 1:A newspaper should be lively and should avoid timeworn stuff。

隐写例句 2:A newspaper should be lively and should avoid trite stuff。

原始例句为百度百科讲解单词“hackneyed”含义时使用的例句。“hackneyed”所在替换集为 {timeworn, hackneyed, trite},因此通过同义词替换嵌入信息生成的隐写例句可能为上述隐写例句 1 和例句 2。3 个例句中出现的 3 个同义词具有相同的上下文单词。本文利用 Word2vec 获得上下文和这 3 个同义词的词向量后,根据式(2)、式(3)计算的同义词“hackneyed”在当前上下文中的合适度约为 1.244 3,而“timeworn”和“trite”的合适度分别为 0.182 3和 1.186 2。由此可见,在原始例句中,原始使用的同义词确实具有最大的上下文合适度,而经过同义词替换后的隐写例句中,同义词的上下文合适度要比替换前的小。“hackneyed”“timeworn”和“trite”的相对词频分别约为 0.452 632、0.073 684 2 和0.473 684,即“trite”比“hackneyed”具有更高的使用频率。如果仅从词频特性来提取特征,则很难准确区分原始例句和隐写例句 2。因此,使用词向量能更准确地刻画同义词的上下文合适度,从而提高隐写分析特征检测同义词替换隐写文本的能力。

一方面,如果句子中使用的某个同义词的上下文合适度不是其替换集中所有元素的上下文合适度的最大值,那么该词是为了嵌入信息将原始的词进行替换。另一方面,如果一个同义词的上下文合适

度与其替换集中所有元素的上下文合适度的最大值差距越大,那么在当前上下文中正常使用该词的概率越低,该词是替换过的概率越高。从这两方面考虑,本文提取 2 个用于区分基于同义词替换的隐写文本和正常文本的隐写分析特征。

假设待分析文本中出现的所有同义词为 $s_1,s_2,\cdots,s_n$,依次计算每个同义词 $s_i(1\leqslant i\leqslant n)$ 所在上下文中的合适度并表示为 F_i。将每个同义词 s_i 依次替换为其所在替换集中的可替换元素,分别计算每个可替换元素在当前上下文中的合适度,再将该替换集所有元素对应的合适度值进行对比,选出最大值并记为 $F_i^{\max}$。在此基础上,计算 2 个检测特征 λ 和 θ,计算公式如下:

$$\lambda=\frac{1}{n}\sum_{i=1}^{n}[F_i=F_i^{\max}] \tag{4}$$

$$\theta=\frac{1}{n}\sum_{i=1}^{n}(F_i-F_i^{\max})^2 \tag{5}$$

$$[F_i=F_i^{\max}]=\begin{cases}1,F_i=F_i^{\max}\\0,F_i\neq F_i^{\max}\end{cases} \tag{6}$$

其中,λ 表示文本中出现的可替换元素(同义词)的上下文合适度为其对应替换集中上下文合适度最大值所占的比例。通常正常文本中 λ 的值较大,趋近于 1。在隐写文本中,由于一个同义词的编码值与待嵌入信息的编码值不一致时,为了嵌入信息该同义词将被其他词替换;当两者一致时,保持不变。因此,与嵌入信息前相比,λ 值将会大幅降低。特别是当所有同义词都被嵌入信息即满嵌时,会导致将近一半的具有最大合适度的同义词被替换,此时 λ 的值将趋近于 0.5。θ 表示文本中每个可替换元素的上下文合适度与其对应替换集中上下文合适度最大值之差的平均平方值。通常正常文本中的 θ 值比较小,而隐写文本中 θ 值相对较大。由此可知,λ 和 θ 可有效区分基于同义词替换的隐写文本和正常文本。

1.4 算法描述

通过上述过程,本文将从每个文本中提取出2 个隐写分析特征,然后利用贝叶斯分类器进行训练和检测,从正常文本中识别出基于同义词替换的隐写文本。

1)训练过程

基于 Word2vec 的隐写分析算法的训练过程具体如下:

步骤 1 Word2vec 采用 Skip-gram 模型训练语料库获得字典中所有词的词向量。

步骤 2 利用可替换同义词构建同义词库。

步骤 3 对于训练集中的每一个训练文本进行遍历并利用同义词库检索出文本中出现的所有同义词,该同义词序列记为 $S=\{s_1,s_2,\cdots,s_n\}$,n 为文本中出现的可替换同义词的个数。记录每个同义词所在的上下文,第 i 个同义词 s_i 的上下文为 $c(s_i)=\{w_{i,1},w_{i,2},\cdots,w_{i,2k}\}$,其中 $1\leqslant i\leqslant n$。当上下文窗口的大小不满 k 时用零向量进行填充。

步骤 4 遍历同义词序列 S,检索同义词库并记录每个同义词 s_i 对应的替换集 $SS(s_i)=\{ss_{i,1},ss_{i,2},\cdots,ss_{i,x}\}$。

步骤 5 利用已训练好的词向量库将上述步骤中检索到的同义词、同义词替换集、上下文单词转换为相应的词向量。

步骤 6 根据式(2)依次计算同义词 s_i 与上下文 $c(s_i)$ 中每个单词的相关度。

步骤 7 根据式(3)计算 s_i 在上下文 $c(s_i)$ 下的上下文合适度 F_j。

步骤 8 按照计算 F_i 的方式,计算 s_i 对应替换集 $SS(s_i)$ 中替换元素 $ss_{i,j}$ 与上下文 $c(s_i)$ 中各单词的相关度,$1\leqslant j\leqslant x$,并进一步计算 $ss_{i,j}$ 在上下文 $c(s_i)$ 下的合适度,记为 $F'_{i,j}$,同时计算 $F_i^{\max}=\max(F'_{i,0},F'_{i,1},\cdots,F'_{i,j},\cdots,F'_{i,x})$。

步骤 9 根据式(4)和式(5),计算隐写分析特征 λ 和 θ,构成特征向量。若对应的样本是隐写文本,则给定特征向量的类别标签为 1;若是正常文本,则给定类别标签为 -1。

步骤 10 重复步骤 3 ~ 步骤 9 得到所有训练文本的特征向量。

步骤 11 将所得特征向量输入到贝叶斯分类模型中进行训练得到分类器。

2)检测过程

基于 Word2vec 的隐写分析算法的检测过程具体如下:

步骤 1 对于每个测试文本,重复训练过程中的步骤 3 ~ 步骤 9 提取对应的 2 个隐写分析特征。

步骤 2 将隐写分析特征输入到训练过程中步骤 11 所训练好的贝叶斯分类器进行检测,输出待检测文本的类别。若输出结果为 1,则当前测试文本为隐写文本;否则为正常文本。

2 实验结果与分析

2.1 实验设置

本文实验利用 Word2vec 工具在 Google News 语料库上进行训练获得一个公开词向量库。该词向量库是利用 Skip-gram 模型在约 1 000 亿字的语料库上训练所得,总共包含约 3 000 000 个不同单词的

词向量,每个词向量的维数为 300。从 Gutenberg 语料库中随机下载 5 000 篇世界文学名著组成原始文本集。对于每个原始文本,利用 T-lex 隐写工具按照 100%、75%、50%、25% 这 4 种嵌入率分别嵌入随机生成的秘密信息,生成相应的隐写文本。文献[5]提出的基于矩阵编码的隐写方法(MC)提高了抗隐写分析检测的能力。为验证本文方法的性能,使用 MC 以约 42.8% 的嵌入率(即采用(7,3)的矩阵编码)生成 5 000 个隐写文本。从原始文本集中选取 3 000 个正常文本,以及每种嵌入率的隐写文本中选取 2 000 个隐写文本组成训练集,利用贝叶斯模型训练分类器,剩余的样本组成测试集。

2.2　特征分析

为直观形象地验证隐写分析特征的有效性,本文随机选取 100 个原始文本和 100 个利用 T-lex 以 100% 嵌入率生成的隐写文本,通过本文方法分别提取隐写分析特征 λ 和 θ,结果如图 2、图 3 所示。可以看出,原始文本中的 λ 值集中分布在 0.85 附近,而隐写文本中的 λ 值集中分布在 0.5 附近,两者之间的差距明显,易于区分。尽管隐写文本中的 θ 值比较分散,但远大于原始文本中的 θ 值。由此可见,2 个特征均具有较大区分度,能较好地区分这 2 类不同的文本,说明了本文方法能有效地检测基于同义词替换的隐写文本。

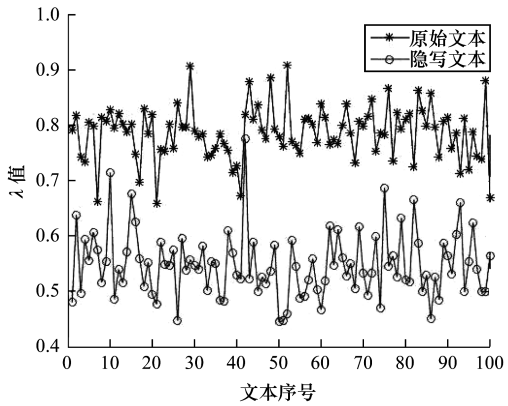

图 2　部分原始文本和隐写文本的 λ 值对比结果

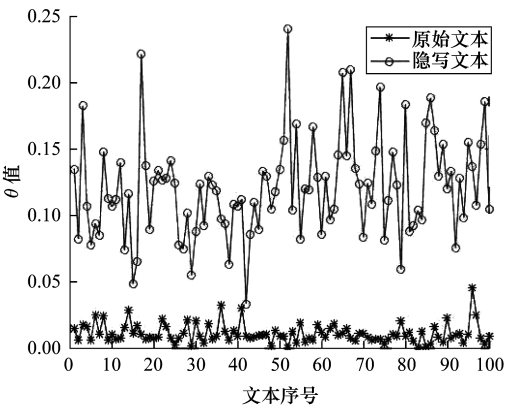

图 3　部分原始文本和隐写文本的 θ 值对比结果

2.3　与其他方法的隐写分析性能对比

为准确地评估隐写分析算法的可靠性,本文通过精确率(P)和召回率(R)表示算法检测结果。

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

其中,TP 表示隐写文本正确判定为隐写文本的数量,FP 表示正常文本错误判定为隐写文本的数量,FN 表示隐写文本错误判定为正常文本的数量。分别使用本文方法、文献[10]方法(NRF)和文献[11]方法(PP)对 T-lex、MC 生成的隐写文本进行检测,结果见表 1。可以看出,本文方法对不同隐写算法和嵌入率生成的隐写文本都有较好的检测性能,且优于同类的隐写方法。

表 1　不同隐写分析方法检测性能对比

隐写文本	性能指标	PP 方法	NRF 方法	本文方法
T-lex-25%	TP	2 076	2 306	2 591
	$TP + FP$	2 436	2 539	2 801
	$TP + FN$	3 000	3 000	3 000
	精确率/%	85.22	90.82	92.50
	召回率/%	69.20	76.86	86.36
T-lex-50%	TP	2 533	2 837	2 867
	$TP + FP$	2 691	2 906	2 905
	$TP + FN$	3 000	3 000	3 000
	精确率/%	94.12	97.62	98.69
	召回率/%	84.43	94.56	95.56
T-lex-75%	TP	2 776	2 940	2 942
	$TP + FP$	2 843	2 969	2 946
	$TP + FN$	3 000	3 000	3 000
	精确率/%	97.64	99.02	99.86
	召回率/%	92.53	98.00	98.06
T-lex-100%	TP	2 896	2 960	2 962
	$TP + FP$	2 941	2 974	2 962
	$TP + FN$	3 000	3 000	3 000
	精确率/%	98.46	99.52	100.00
	召回率/%	96.53	98.66	98.73
MC	TP	2 127	2 221	2 535
	$TP + FP$	2 224	2 353	2 600
	$TP + FN$	3 000	3 000	3 000
	精确率/%	95.63	94.39	97.50
	召回率/%	70.90	74.03	84.50
总计	平均精确率/%	94.21	96.27	97.71
	平均召回率/%	82.72	88.42	92.64

在嵌入率相同的情况下,本文方法具有比其余 2 种方法更高的精确度和召回率,嵌入率越高,检测精确度和召回率随之提高。当嵌入率超过 50% 时,

所有隐写分析方法对隐写文本的检测能力均很强，此时嵌入率对检测性能的影响不明显，但本文方法仍具有比其他方法稍高的精确率和召回率，特别是当嵌入率为 100% 时，基本能完全区分隐写文本和正常文本，精确率达到 100%，召回率达到 98.73%。当嵌入率较低时，PP、NRF 方法的检测性能明显低于本文方法。如对于嵌入率为 25% 的 T-lex 隐写文本，PP、NRF 方法的召回率均低于 80%，而本文方法的召回率达到 86.36%，对应的精确率达到 92.5%，高于 PP、NRF 方法的召回率。对于通过 MC 生成的隐写文本，PP、NRF 方法在检测精确率上与本文方法的差距不大，精确率依次为 95.63%、94.39%、97.5%，但 PP、NRF 方法的召回率分别为 70.9% 和 74.03%，远低于本文方法的召回率 84.5%。由此可见，对于低嵌入率隐写文本的检测，本文方法比 PP、NRF 方法更有优势，具有更好的检测性能。另外，表 1 中给出了 PP、NRF、本文方法的平均精确率和召回率分别为 94.21%、82.72%、96.27%、88.42%、97.71%、92.64%，可见，本文方法具有较好的综合检测性能。

3 结束语

本文从词的分布式表示出发，提出一种基于 Word2vec 的针对同义词替换隐写的隐写分析方法。该方法利用 Word2vec 工具获得相关单词的词向量后，利用词向量包含的丰富语义信息度量一个同义词在特定上下文中的合适度，并从中提取检测特征用来区分基于同义词替换的隐写文本和正常文本。实验结果表明，本文方法可以有效检测低嵌入率时基于同义替换的隐写文本，获得了比同类方法更好的检测性能。但本文方法设计的词相关度、合适度估算模型相对简单，下一步将考虑用词的词性、搭配、句法结构等度量同义词的上下文合适度，提取更有效的特征，以提高低嵌入率时隐写分析方法的检测性能。

参考文献

[1] WILSON A, KER A D. Avoiding detection on Twitter: embedding strategies for linguistic steganography[J]. Electronic Imaging, 2016(8): 1-9.

[2] HU H, ZUO X, ZHANG W, et al. Adaptive text steganography by exploring statistical and linguistical distortion[C]//Proceedings of the 2nd International Conference on Data Science in Cyberspace. Washington D.C., USA: IEEE Press, 2017: 145-150.

[3] WINSTEIN K. Lexical steganography through adaptive modulation of the word choice hash[EB/OL]. [2018-01-07]. http://alumni.imsa.edu/~keithw/tlex/lsteg.ps.

[4] 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.

[5] 杨潇，李峰，向凌云. 基于矩阵编码的同义词替换隐写算法[J]. 小型微型计算机系统，2015，36(6)：1296-1300.

[6] 霍林，肖豫川. 基于二元依存同义词替换隐写算法[J]. 计算机应用研究，2018，35(4)：1174-1178.

[7] 罗纲，孙星明，向凌云，等. 针对同义词替换信息隐藏的检测方法研究[J]. 计算机研究与发展，2008，45(10)：1696-1703.

[8] YU Z, HUANG L, CHEN Z, et al. Steganalysis of synonym-substitution based natural language watermarking[J]. International Journal of Multimedia and Ubiquitous Engineering, 2012, 4: 21-34.

[9] CHEN Z, HUANG L, MIAO H, et al. Steganalysis against substitution-based linguistic steganography based on context clusters[J]. Computers and Electrical Engineering, 2011, 37(6): 1071-1081.

[10] CHEN Z, HUANG L, YANG W. Detection of substitution-based linguistic steganography by relative frequency analysis[J]. Digital Investigation, 2011, 8(1): 68-77.

[11] XIANG L, SUN X, LUO G, et al. Linguistic steganalysis using the features derived from synonym frequency[J]. Multimedia Tools and Applications, 2014, 71(3): 1893-1911.

[12] HINTON G E. Learning distributed representations of concepts[C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society. Amherst, USA: Erlbaum Associates, Inc., 1986: 12.

[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-01-07]. https://arxiv.org/abs/1301.3781.

[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of NIPS'13. [S.l.]: Curran Associates, Inc., 2013: 3111-3119.

[15] 梁军，柴玉梅，原慧斌，等. 基于深度学习的微博情感分析[J]. 中文信息学报，2014，28(5)：155-161.

[16] 孙紫阳，顾君忠，杨静. 基于深度学习的中文实体关系抽取方法[J]. 计算机工程，2018，44(9)：164-170.

编辑 陆燕菲