针对文本水印摘要攻击的语义损失量评估方法*

张雅轩, 何 路, 房鼎益
(西北大学 信息科学与技术学院, 西安 710127)

摘 要: 为了控制文本水印自动摘要攻击造成的语义信息损失,在已有的自动摘要评估方法的基础上,针对自动摘要文本水印攻击算法提出一种评估文本语义损失度的算法。该方法通过量化句子语义,合理计算摘要攻击造成的语义损失;并分析了攻击造成语义损失的主要因素,以及这些因素和语义损失量之间的数学关系。该方法能够从语义信息损失的角度更真实地评测摘要攻击算法的失真度,实现了自动化评估。

关键词: 语义损失量; 摘要攻击; 文本水印; 自动摘要
中图分类号: TP309 **文献标志码**: A **文章编号**: 1001-3695(2012)09-3483-03
doi:10.3969/j.issn.1001-3695.2012.09.075

Method of meaning loss evaluation for summary attack in text watermarking

ZHANG Ya-xuan, HE Lu, FANG Ding-yi
(*School of Information & Technology*, *Northwest University*, *Xi' an* 710127, *China*)

Abstract: Summary attack on natural language watermarking was likely to cause excessive loss of the texts' original meaning. In order to control the loss, this paper proposed a method based on the existing evaluation method of automatic summarization. It quantized the meaning of every sentences, then evaluated the meaning loss which caused by the summary attack. It analyzed the major factors of meaning loss, as well as the relationship between these factors and the meaning loss. This method gave the attack algorithm's distortion a more realistic evaluation and was an automatic evaluation method.

Key words: meaning loss; summary attack; text watermarking; automatic summarization

数字水印是信息隐藏技术的一个重要研究方向。基于自然语言处理的文本水印技术在保护可公开传播信息的版权及保证秘密和隐私信息的安全等方面具有重要的研究意义和广泛的应用前景。而相应的攻击方法研究则是促进文本水印技术发展所必不可少的,两者相辅相成。

一次成功的攻击不仅需要破坏载体文本所嵌入的秘密信息,而且不能破坏载体的使用价值,要保证攻击后的文本阅读流畅,保留文章所要表达的主要信息。如何计算攻击带来的信息损失,如何客观地评价攻击的失真度,是评价攻击算法优劣所必须讨论的问题,对文本水印的研究有着重要意义。

摘要攻击是文本水印攻击方法中的一种,这种攻击方法可以有效地破坏水印嵌入位,但删减文字的同时容易造成较大的语义损失[1]。因此需要一种衡量摘要攻击所造成的语义损失的方法,这就涉及到自动摘要评测方法。现有的自动摘要评测方法主要分为内部评价方法和外部评价方法。内部评价方法通过直接分析文摘的质量来评价文摘系统;外部评价方法是一种间接的评价方法,将文摘应用于某个特殊任务中,根据摘要功能提高任务的效果来评价摘要系统的性能[2,3]。本文所研究的内容显然需要用到的是内部评价方法,然而内部评价方法需要有理想文摘进行比较。理想文摘一般由专家给出,不易获得,且主观性强。

对于摘要攻击的语义损失评估,人们更希望找到摘要攻击方法和语义损失量之间的关系,从而实现不需要专家文摘作对比,并可以对任意摘要系统攻击后的载体计算语义损失量,从语义信息损失的角度评价自动摘要攻击的失真度。

本文提出一种针对自动摘要攻击的语义损失量评估方法,量化全文语句的语义信息,合理计算摘要攻击造成的语义损失。

1 相关理论分析

1.1 自动摘要技术

自动文摘是利用计算机自动地从文本或文本集合中提炼出能准确、全面反映文本主要内容的精简、连贯的短文,以满足用户的需求。

自动文摘主要采用以下四种方法[4]:自动摘录(基于统计的自动文摘)、基于理解的自动文摘、信息抽取和基于结构的自动文摘。本文用到的自动文摘系统 MEAD 就是使用基于统计的自动文摘方法。

基于统计的自动文摘将文本视为句子的线性序列,将句子视为词的线性序列。它通常分四步进行:a)计算词的权值;b)计算句子的权值;c)对原文中的所有句子按权值高低降序排列,权值最高的若干句子被确定为文摘句;d)将所有文摘句按照它们在原文中的出现顺序输出。

为了估计摘要删掉的句子对原文语义所造成的损失,本文所提出的算法需要量化地表示每个句子在文章中所占的语义

收稿日期: 2012-01-15; **修回日期**: 2012-03-06 **基金项目**: 陕西省教育厅 2009 年度科学研究计划资助项目(09JK751)

作者简介:张雅轩(1987-),女,陕西西安人,硕士,主要研究方向为自然语言信息隐藏技术(zhangyaxuan310@163.com);何路,男,讲师,博士研究生,主要研究方向为网络与信息安全、数字内容安全、自然语言信息隐藏;房鼎益,男,教授,博导,主要研究方向为网络与信息安全、分布计算系统和软件安全与保护等.

信息比重。本文选择使用 MEAD 作为量化句子语义的工具。

MEAD[6] 是一个多语言摘要工具。MEAD 在进行自动摘要前会算出原文每个句子的三个特征值:centroid、length 和 position。依据这三个值给句子打分,得到句子的权值,据此生成摘要。在这三个特征值中,centroid 值反映了一个句子在文章中的重要程度。在 MEAD 中,centroid 是指文章中所有 count * IDF 值大于一个阈值的词汇的集合[5]。Centroid 特征值是一个句子对文章 centroid 向量的余弦覆盖,这个值可以说明句子反映文章主题的程度。本文提取该值作为句子的语义量,实现文章语义的量化。

1.2 摘要攻击

在不窜改载密文本原义的情况下,增加或删除句子都会破坏水印比特同步。He 等人[1] 提出摘要攻击的方法,利用自动摘要工具,按照给定的攻击率(摘掉的词数与原文词数之比),逐段删掉载密文本中不太重要的句子,保留大部分重要信息,破坏水印比特同步。这种攻击方法可以有效地破坏水印信息,但删减句子的同时也容易破坏文章原义,造成原文所要表达的信息流失。因此,需要对摘要攻击后的文章做语义损失评估。

本文所用的摘要攻击工具,是西北大学网络与安全实验室的自动攻击平台[7],该平台实现了摘要攻击方法。用户可以给定攻击力度和压缩比,逐段对文章进行摘要攻击。所谓攻击力度是指摘要后去掉的水印嵌入单元个数占原文总嵌入单元个数的百分比。如攻击力度设为 10,则攻击后应去掉 10% 的嵌入单元。所谓压缩比是指摘要去掉句子的比例,本文所用的攻击方法是按段攻击,那么如果压缩比是 0.8,也就是说要保留该段 80% 的句子,去掉 20%。攻击所用的工具是 Microsoft Word 自带的摘要工具。

2 摘要攻击语义损失量评估算法

文章中每个句子都包含有一定的语义信息,传达给读者这篇文章所要表达的内容。有些句子涉及文章主旨,有些句子只是修饰辅助。摘要攻击按照某种既定的方法摘掉文中的某些句子,因此需要评价摘要攻击所摘掉的句子造成多少语义信息的流失。本文通过量化句子语义值的方法,累积摘掉句子的语义值,计算摘要的语义损失量,并针对摘要攻击方法进行语义损失量评估实验。通过数据分析得到攻击和语义损失量之间的关系。实验包括机器计算句子语义值和人工给出句子语义值两组数据。两组数据计算得到的语义损失量的差值可对摘要攻击语义损失量进行校正,使结果更接近人对摘要语义损失的评价。事实上这种分析方法也可以用于其他摘要攻击方法。

2.1 语义损失量算法

语义损失量是指一篇摘要相对于原文所丢失的语义信息的程度。如何合理、量化地评价摘要的语义损失量,本文提出的语义损失量算法描述如下:

a)分句。用 MEAD 给原文分句,生成特定格式的文本文件。

b)计算句子语义值。

c)标准化语义值。标准化步骤 b)所得的语义值,文章所有句子语义值总和为 1。

d)计算文章所有句子两两之间的句子相似度。

e)聚类。利用步骤 d)的结果,按层次聚类法,将相似度高的句子聚成一类。

f)计算语义损失量。对比原文和摘要,找出被摘掉的句子。按被摘掉句子的语义值在其分类语义总值所占比例缩小该句语义值。所有被摘掉句子的语义值之和即为该篇摘要的语义损失量。

下面对各个步骤作详细解释。

步骤 b)中计算语义值的方法分两种:机器打分和人工打分。对于机器计算语义值,本文利用 MEAD 给完成分句的文章计算每句的 centroid 特征值,以此作为句子的语义值。机器并不能很好地判断每句话在文中所占的语义比重,为了得到更准确的语义损失量,本文采用人工打分的方式标注句子语义,校正机器计算的结果。具体方法在第 3 章的实验部分会具体阐述。

步骤 c)中的标准化语义值的方法是设 C_i 为第 i 个句子的 centroid 值,那么第 i 句标准化的语义值为

$$\hat{C}_i = \frac{C_i}{\sum_{j=1}^{n} C_j} \quad i = 1,2,\cdots,n \tag{1}$$

其中 n 为文章句子总数。标准化后,整篇文章句子语义值之和为 1,即

$$\sum_{i=1}^{n} \hat{C}_i = 1 \quad i = 1,2,\cdots,n \tag{2}$$

步骤 d)中计算句子相似度的方法有很多种,本文利用了 MEAD 提供的计算句子相似度的方法,将 MEAD 中计算文章首句和其他句子间相似度的 simwithfirst. pl 文件改写成计算所有句子两两之间的相似度。

步骤 e)中利用上一步的句子相似度计算结果,用层次聚类法将语义相近的句子划分为一类。

步骤 f)中计算语义损失量的具体方法如下:

首先,对比原文和摘要,找出所有被摘掉的句子。设被摘掉的句子 S_i 所在的分类 $S = \{S_1, S_2, \cdots, S_m\}$,$S_i$ 标准化后的语义值为$\hat{C}_i$,那么摘掉句子 S_i 的语义损失量为

$$L_i = \hat{C}_i \times \frac{\hat{C}_i}{\sum_{j=1}^{n} \hat{C}_j} \quad j = 1,2,\cdots,m \tag{3}$$

所有被摘掉句子的语义损失量之和,即为这篇摘要的语义损失量:

$$L = \sum_{i=1}^{n} L_i \quad i = 1,2,\cdots,n \tag{4}$$

标准化被摘掉句子的语义值的意义在于语义相近的句子包含了重复的语义信息,去掉该句并不会造成很大的语义损失,所以按比例缩小该句语义值能够更合理地得到去掉该句后文章的语义损失量。

这种语义损失量计算方法能量化地表示一篇摘要的语义损失,并考虑到相似句子包含语义信息重复的情况。合理缩减摘掉该句所造成的语义损失量,能够更为真实地反映摘要相对于原文所损失的语义信息量。

2.2 摘要攻击语义损失评估算法

仅由机器量化句子语义值,未必符合读者对文章的理解。因此本文提出人工量化句子语义值,计算语义损失量,用这个结果校正机器语义值所得的语义损失量。这样计算的语义损失量会更符合人的认识。

对于不同的摘要攻击方法,影响摘要语义损失的因素是不

同的。但对于同一种攻击方法,无论是机器量化句子语义值,还是人工量化,影响摘要语义损失的因素是相同的。可以用数学方法找出这些因素和语义损失量之间的关系。

本文所采用的攻击方法是按照给定的攻击力度和压缩比,按段落进行摘要,破坏足量的嵌入单元。

摘要攻击语义损失评估算法描述如下:

a)用机器给文章句子量化语义值。按 2.1 节所述方法计算摘要的语义损失量。

b)人工给文章句子量化语义值。计算摘要的语义损失量。

c)分析攻击造成语义损失的因素,用方差分析的方法找出主要因素。设主因子 $X=\{x_1,x_2,\cdots,x_n\}$。

d)用机器语义值和人工语义值所得的两组数据,分别计算主因子和语义损失量的回归方程。设机器语义值所得回归方程为 $F(X_1)$,人工语义值所得回归方程为 $F(X_2)$。

e)给定攻击因子的值,对任意一篇文章进行摘要攻击。用机器量化摘要中句子的语义值,计算摘要的语义损失量 L。将因子的值代入回归方程,最终校正得到的语义损失量为

$$\mathrm{LOSE}=L+F(X_2)-F(X_1) \tag{5}$$

对于本文所采用的攻击方法,影响摘要语义损失量的主要因素就是攻击力度和压缩比。

3　实验及数据分析

本文的实验针对 1.2 节所述文本水印摘要攻击方法,用 182 篇文本样本进行语义损失评估实验。实验分别对十种攻击力度和九种压缩比的摘要攻击进行语义损失量计算。攻击力度分别为 2,4,…,20,压缩比分别为 0.1,0.2,…,0.9。

3.1　机器标注语义值评估实验

首先,对所有摘要攻击结果用 MEAD 标注语义值的结果按照第 2 章中的方法计算语义损失量。

实验表明,语义损失量随着攻击力度的增加而增加,随着压缩比的增加而减少。这说明摘要攻击越分散越均匀,语义损失量就会越小。图 1 直观地反映了这一趋势。

用二因子方差分析检验攻击力度和压缩比这两个因子对语义损失量的影响分析结果如表 1 所示,攻击力度因子和压缩比因子的 P 值为 0,远远低于 0.01,这说明攻击力度和压缩比对语义损失量的影响是显著的。

表 1　方差分析表

来源	自由度	*SS*	*MS*	*F*	*P*
攻击力度	9	0.031 718 0	0.003 524 2	37.24	0.000
压缩比	8	0.013 154 6	0.001 644 3	17.38	0.000
误差	72	0.006 813 6	0.000 094 6		

3.2　人工标注语义值评估实验

在 182 篇样本中抽取 15 篇文章人工标注语义值。按照句子在文章中的重要程度和信息量大小给每个句子标记语义值。

由于人工标记十分费时费力,所以首先标注了 5 篇文章,用这 5 篇文章的数据进行最小样本量估计。同样用方差分析的方法,将置信度设为 90%,通过 mintab 计算得到最小样本量为 17。当标记了 15 篇之后,再次估计最小样本量为 12,说明这 15 篇文章的数据可以反映样本总体攻击力度和压缩比对语义损失量的影响,置信度为 90%。

图 2 反映了语义损失量随着攻击力度和压缩比变化的趋势,可以看出,这与机器打分的趋势是相同的,即语义损失量随着攻击力度的增加而增加,随着压缩比的增加而减少。

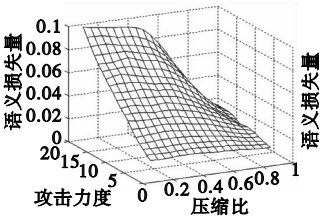

图1　语义损失量与攻击力度和压缩比的关系图(机器标注语义值)

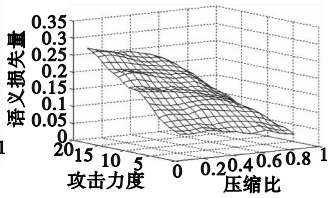

图2　语义损失量与攻击力度和压缩比的关系图(人工标注语义值)

3.3　回归方程

利用以上实验数据,可以求得以攻击力度、压缩比为因子,以语义损失量为响应的回归方程。设攻击力度为 x_1,压缩比为 x_2,语义损失量为 y,则机器标注语义值实验数据分析所得回归方程为

$$y=0.00023x_1^2-0.077270x_2^2-0.005245x_1x_2+0.005391x_1+0.091898x_2-0.014001 \tag{6}$$

人工标注语义值实验数据分析所得回归方程为

$$y=0.00018x_1^2-0.262025x_2^2-0.0115978x_1x_2+0.0178043x_1+0.269515x_2-0.0192853 \tag{7}$$

设给定攻击力度和压缩比的值,机器打分回归方程求得的语义损失量为 y_1,人工打分回归方程求得的语义损失量为 y_2。机器评估和人工评估之间的差值就可用于校正语义损失量,得到一个更近似于人为的评估结果。当设定攻击力度和压缩比,用第 2 章所述评估方法得到一篇摘要攻击后的文本的语义损失量 y_3 的时候,就能够计算出最终校正后的摘要语义损失量:

$$\mathrm{LOSE}=y_3+y_2-y_1 \tag{8}$$

LOSE 是经过校正的语义损失量结果,更真实地反映这篇被攻击的文章相对原文所损失的语义信息。

通过这种方法得到摘要攻击和语义损失之间的数学关系式,在给定攻击力度和压缩比的情况下就可通过机器计算语义损失量得到一个类似人工校正后的更为真实合理的语义损失量计算结果。

4　结束语

本文提出一种利用 MEAD 摘要工具计算摘要攻击后文本的语义损失量的方法,并就该方法针对一种摘要攻击进行语义损失量评估实验,分析攻击力度和压缩比对语义损失量造成的影响,得到三者之间的数学关系式。最终提出如何校正机器计算的语义损失量,得到一个更为合理的评估结果。

事实上,本文真正的意义在于提出这种计算语义损失量的方法。通过这种方法可以量化、合理地计算出摘要攻击所造成的语义损失,更为真实地评价一个摘要攻击算法的失真度。今后,通过更全面深入的研究,将研究更好的量化句子语义值的方法,以及计算句子相似度的方法。还会研究更多不同的摘要攻击方法,找到其他造成语义损失的主要因素,得到更接近真实的语言损失评估结果。

参考文献:

[1] HE Lu, GUI Xiao-lin, JIN Yao-ping. Summary attack on spread spectrum vector natural language watermarking[J]. Journal of Computational Information Systems, 2012,8(4):1-8. (下转第 3488 页)

(上接第 3485 页)

[2] 张培颖. 基于语义相似度的自动文摘评价方法[J]. 计算机工程与应用,2009,45(25):145-147.

[3] 蒋勇铭,陈微熙,张仲贵. 基于句子相似度的自动文摘评价方法[J]. 中国西部科技,2011,10(29):22-23.

[4] 刘挺,王开铸. 自动文摘的四种主要方法[J]. 情报学报,1999, 18(1):10-19.

[5] RADEV D R, JING Hong-yan, BUDZIKOWSKA M. Centroid-based summarization of multiple documents[J]. Information Processing and Management, 2004,40(6):919-938.

[6] RADEV D, BLITZER J, WINKEL A, *et al*. MEAD[EB/OL]. (2003-11-21). http://www.summarization.com/mead/.

[7] 西北大学信息学院数字内容安全研究小组. 自动攻击[EB/OL]. http://dc-security.org/download/autoattacker.html.

[8] ANI H A, NENKOVA A, PASSONNEAU R, *et al*. Automation of summary evaluation by the pyramid method[C]//Proc of Conference of Recent Advances in Natural Language Processing. 2005.

[9] RADEV D, ALLISON T, GOLDENSOHN S B ,*et al*. MEAD:a platform for multidocument multilingual text summarization[C]//Proc of the 4th International Conference on Language Resources and Evaluation. 2004.

[10] MANI T. Automatic summarization[M]. Amsterdam, Philadelphia: John Benjamins Publishing Company, 2001.