文章编号：1001-0920(2009)12-1861-04

基于移位变换的句子层自然语言信息隐藏算法

刘玉玲[1]，孙星明[1]，辛国江[2]
(1.湖南大学 计算机与通信学院，长沙 410082；2.湖南网络工程职业学院 信息技术系，长沙 410002)

摘 要：针对现有的句子层自然语言信息隐藏方法存在的问题，提出一种基于句子层移位变换规则的中文自然语言信息隐藏算法.首先利用汉字数学表达式对中文文本信号数字化；然后通过句子的移位变换改变句子词序以嵌入秘密信息.实验结果和分析表明，该算法实现简单、编码容量较大，同时秘密信息隐藏在自然语言文本句子层词序变换中，不改变文本语法、语义和风格，具有较好的隐蔽性.

关键词：信息隐藏；自然语言；移位变换；文本数字化
中图分类号：TP391.1 **文献标识码**：A

Algorithm of natural language information hiding based on shift conversion in sentence level

LIU Yu-ling[1], *SUN Xing-ming*[1], *XIN Guo-jiang*[2]
(1. School of Computer and Communication, Hu'nan University, Changsha 410082, China; 2. Department of Information and Technology, Hu'nan Network Engineering Vocational College, Changsha 410002, China. Correspondent: LIU Yu-ling, E-mail: yuling_liu@126.com)

Abstract: An algorithm based on shift conversion is proposed combining with the characteristic of Chinese grammar. A method based on chinese mathematical expression is presented by converting Chinese text into bit string. Then, shift conversion rules are selected by utilizing transform grammar based on case grammar that Chinese philologist proposed. The secret information is embedded by modifying the order of words in the sentence according to the shift conversion rules. The experimental results show that the method can achieve a degree of information-carrying capacity and a better result with the imperceptibility.

Key words: Information hiding; Natural language; Shift conversion; Text digitalization

1 引 言

自然语言信息隐藏是利用自然语言处理技术，将信息隐藏于文本内容中的一种信息隐藏方法，能有效抵御重新排版和 OCR 攻击，受到国内外众多学者的广泛关注.针对基于词汇替换的方法存在隐蔽性不好等问题，有学者提出在句子层隐藏信息[1].句子层自然语言信息隐藏方法利用等价的句法变换，修改句子结构而不改变句子的意义以隐藏信息.目前的研究主要集中在英语，也有学者研究土耳其语.爱尔兰的 Murphy 等[2,3]针对英语的特点，提出了多种保持语义和风格的句法变换.土耳其的 Meral 等[4]针对土耳其语的特点，提出了 21 种适应于土耳其语的句法变换.同时，为了提高容量和增强安全性，句子层自然语言信息隐藏方法需将文本信号进行数字化描述.美国普渡大学的 Atallah 等[5]对文本进行句法分析，得到了相应的句法分析树并进行编码的数字化方法.澳大利亚的 Gupta 等[6]提出了基于句子长度编码的数字化方法.

汉语不同于英语和土耳其语，是一种孤立语.针对现有句法变换主要适用于英文自然语言信息隐藏，而对中文文本不实用等问题，本文提出一种基于移位变换的句子层中文自然语言信息隐藏算法.该算法根据汉语语言学者总结的移位变换规则，利用中文信息处理成果(如分词、词性标注等)，通过移位变换改变句子的词序以隐藏信息；同时，针对现有文本信号数字化方法复杂、编码容量小等问题，提出一

收稿日期：2009-02-10；**修回日期**：2009-04-03.
基金项目：国家 973 计划项目(2006CB303000)；国家自然科学基金重点项目(60736016)；国家自然科学基金面上项目(60873198)；湖南省科技计划项目(2008FJ4221).
作者简介：刘玉玲(1980—)，女，湖南宁乡人，讲师，博士，从事信息隐藏、自然语言处理的研究；孙星明(1963—)，男，湖南益阳人，教授，博士生导师，从事网络信息安全、自然语言处理等研究.

种基于汉字数学表达式的文本信号数字化方法.

2 基于汉字数学表达式的文本数字化方法

句子层自然语言信息隐藏中直接对等价句型进行编码的方法,其编码容量有限,安全性和鲁棒性有待增强.本文提出一种基于汉字数学表达式思想[7]的中文文本信号数字化方法.下面以一个汉字为例,数字化过程包括以下3个转换步骤:

Step1: 令 c 表示一个汉字,用 math(c) 表示与 c 对应的数学表达式,如 c = 在,math(c) = 498lu28.其中:数字为基本部件号,lu 为运算符.汉字数学表达方法的详细内容见文献[8].

Step2: 当 c 为基本部件时,直接转 Step3;当 c 为非基本部件时,比较数学表达式 math(c) 中运算符的优先级,将最低优先级的运算符 op 作为二叉树的根.op 左边的基本部件和运算符即为左子树,用 l 表示;op 右边的基本部件和运算符即为右子树,用 r 表示.这样,math(c) 可以写成 l op r 的二叉树形式,如图1所示.

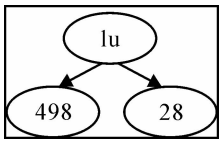

图1 汉字"在"的二叉树示例

Step3: 当 c 为基本部件时,直接用0表示;当 c 为非基本部件时,由 Step2 得到二叉树形式 l op r.令 sum(l) 表示 l 中所有叶子节点对应的部件号之和,sum(r) 表示 r 中所有叶子节点对应的部件号之和.利用下式进行判断:

$$\text{mark}(c)=\begin{cases}0, & \text{sum}(l)<\text{sum}(r);\\ 1, & \text{sum}(l)\geqslant\text{sum}(r).\end{cases}\tag{1}$$

即可将汉字 c 的值 mark(c) 编码为0或1.

3 句子层移位变换规则

首先引入如下符号:N 表示名词性成分,V 表示动词,$N_{时}$ 表示时间名词,$N_{工具}$ 表示工具名词,$N_{处}$ 表示处所名词,⇔ 表示左右两式可以进行等价的句法变换.

汉语语言学者以汉语同义句的比较为基础,根据同义句中的同义成分进行移位变换,为汉语单句设计了5类移位变换规则,即介词格变换、受事格变换、状态存在句变换、非控格变换及异形句变换[9],这为基于句法变换的句子层中文自然语言信息隐藏方法提供了有利基础.针对文献[9]总结的6种介词格的24条移位变换规则,本文最终选择了如下6种移位变换规则:

1) $N_{施}$ + 在 · $N_{时}$ + V ⇔ 在 · $N_{时}$ + $N_{施}$ + V;

2) $N_{施}$ + 对 · N + V ⇔ 对 · N + $N_{施}$ + V;

3) $N_{施}$ + 为 · V_2 + V_1 ⇔ 为 · V_2 + $N_{施}$ + V_1;

4) $N_{施}$ + 从 · $N_{时}$ + V ⇔ 从 · $N_{时}$ + $N_{施}$ + V;

5) $N_{施}$ + 向 · $N_{处}$ + V ⇔ $N_{施}$ + V + 向 · $N_{处}$;

6) $N_{施}$ + 用 · $N_{工具}$ + V ⇔ 用 · $N_{工具}$ + $N_{施}$ + V.

4 隐藏算法与提取算法

为简便起见,令 $T=T_1T_2\cdots T_{|T|}$ 为原始文本,T' 表示隐藏信息后的文本,T_i 表示文中的一个句子,$|T|$ 表示 T 的句子数.令待隐藏的秘密信息为 $M=m_0m_1\cdots m_{|M|}$,$m_j\in\{0,1\}$ 表示 M 的第 j 位,$|M|$ 表示 M 的长度.令 BIN(T_i) 表示对句子 T_i 进行数字化的方法,$S_i=s_{i0}s_{i1}\cdots s_{ij}\cdots$ 表示数字化后的序列,$s_{ij}\in\{0,1\}$ 表示 S_i 中的第 j 位,$|S_i|$ 表示 S_i 的长度.令密钥为 k,$\alpha=k\%|S_i|$ 表示 S_i 中嵌入位置.令第3节中选取的6种介词(词性为 p) 表示为 J = {在,从,对,用,为,向},选取的移位变换规则为 Rule = $\{R_1,R_2,\cdots,R_6\}$.令 Seg(T_i) = $\{w_1/t_1,\cdots,w_e/t_e,\cdots\}$ 表示对 T_i 进行分词及词性标注,w_e 为句中的一个词,t_e 为 w_e 的词性.A = CParse(T_i) 表示对 T_i 进行格语法分析后得到句法规则 A,其中包含句子 T_i 的动词 V 及其施事格和介词格 $N_{施}$,$N_{介}$.Transform(T_i) 表示对句子 T_i 中的介词格按照 Rule 中的规则进行移位变换.Edit(T_i) 表示对 T_i 中的介词进行同义词替换或隐性格变换,其中同义词替换与隐性格变换规则见文献[9].

算法1 隐藏算法

Input: T,M,k;

Output: T'.

1) 初始化:$i=0$,$j=0$;

2) while ($i<|T|$ && $j<|M|$){

// 文本足够长,循环执行步骤3) ~ 11)

3) Seg(T_i) = $\{w_1/t_1,\cdots,w_e/t_e,\cdots\}$;

4) If($\exists w_e\in J$ && $t_e=p$),

Then S_i = BIN(T_i), $\alpha=k\%|S_i|$;

// 对 J 中介词所在句 T_i 进行数字化并确定嵌入位置

5) If($s_\alpha=m_j$), Then $T_i'=T_i$, j++, i++;

// 不需要进行变换即满足嵌入要求,跳出本次循环

6) Else A = CParse(T_i);

7) If($A\notin$ Rule),

Then T_i' = Edit(T_i), i++;

// 当 T_i 不属于可变换规则时,进行同义词替换或隐性格变换,并跳出本次循环

8) Else T_i'' = Transform(T_i),

S_i'' = BIN(T_i'');

// 根据移位变换规则变换后进行数字化

9) If ($s''_{ia} = m_j$),

Then $T'_i = T''_i$, $j++$, $i++$;

10) Else $T'_i = \mathrm{Edit}(T''_i)$, $i++$;

11) Else $i++$;

// 当 T_i 中不存在 J 中介词时，跳出本次循环}

12) Output $T' = T'_1 T'_2 \cdots T'_{|M|} \cdots$

算法 2　提取算法

Input: T', k, $|M|$;

Output: M.

1) 初始化: $i = 0$, $j = 0$;

2) while ($j < |M|$){

3) $\mathrm{Seg}(T'_i) = \{w'_1/t'_1, \cdots, w'_e/t'_1, \cdots\}$;

4) If ($\exists w'_e \in J \&\& t'_e = p$),

Then $S'_i = \mathrm{BIN}(T'_i)$, $\alpha = k\% \mid S'_i \mid$,

$m_j = s'_{i\alpha}$, $j++$, $i++$;

5) Else $i++$;}

6) Output $M = m_0 m_1 \cdots m_{|M|}$.

5　实验结果及分析

5.1　实验语料

实验语料取自北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的 1998 年人民日报语料. 这里仅选取 1 月份的免费语料进行实验. 通过对介词格所在语境的观察发现，介词格一般局限在一个句中，极少有横跨两句的情况. 以句号、问号、叹号、分号、冒号的标点符号作为分句标志，选取的语料共有 44634 个分句.

5.2　实验分析

5.2.1　容量分析

实验语料中每种介词格的出现频率与可变换频率的统计数据如表 1 所示. 经统计，平均每句含有介词格的频率为 43%，这为本文选取介词格进行移位变换提供了有利条件. 根据第 3 节选取的移位变换规则，平均每个句子可进行 0.12 种介词格变换.

表 1　实验数据中介词格变换的统计

Case	“在”	“对”	“为”	“从”	“向”	“用”
Sentences of containing preposition case	10047	3375	2549	1905	1215	399
Transformable sentences	3014	948	397	706	434	90
Transformable percent/%	30.8	28.9	15.6	37.1	35.8	22.6

令 $T_i = c_1 c_2 \cdots c_h$ 为一个句子，c_i 为句中的一个字符，该句的字符总数为 h. 基于句法分析树的编码方法，通过对文本中的每个句子进行句法分析后生成一棵节点数为 r 的句法树 R，可编码成 r 位二进制信息[5]. 基于句子长度编码的方法直接统计句子的字符数 h，然后将 h 表示成二进制形式，编码长度为 $\lceil \log_2 h \rceil$ [6]. 而本文的文本信号数字化方法可将每个汉字编码为 1 位二进制信息，编码长度为 h. 一般地，$r < h$ 且 $\lceil \log_2 h \rceil < h$.

5.2.2　隐蔽性分析

由于自动语义分析技术还不成熟，与其他载体的隐蔽性度量方法相比，自然语言信息隐藏方法的隐蔽性度量极具挑战性[7]. 本文主要通过人工评测方法来判断语句是否流畅以及是否存在歧义性. 这里从语料库中随机选择 100 个嵌入了信息的句子. 其中：“在”格 30 个句子，“对”格 20 个，“从”格和“用”格分别为 15 个句子，“为”格和“用”格分别为 10 个句子. 在不告知进行了何种变换的前提下，请实验室的 10 位同学分别对这些句子的流畅性和无歧义性进行判断. 其中：Normal sentences 表示人工判断无歧义且流畅的句子数，Suspicious sentences 表示人工判断有歧义或不流畅的句子数. 定义成功率

$$\mathrm{SR} = \frac{\text{Normal sentences}}{\text{Total sentences}} \times 100\%.$$

由于人工评测不可避免地存在噪声，这里取平均成功率为 95.4%. 与文献[2]的可靠性相比，本文方法比单个变换的结果低 0.4%，比混合变换的结果高 7.5%.

6　结　　论

现有句子层自然语言信息隐藏方法主要针对英文文本，所采用的句法变换方式不适合中文文本. 本文利用现有的汉语语法研究成果，提出了一种基于移位变换的句子层中文自然语言信息隐藏算法. 该算法具有一定的隐藏容量和较好的隐蔽性，能有效地用于中文文本的隐秘通信和版权保护. 同时还提出了一种基于汉字数学表达式思想的文本信号数字化方法，该方法易于实现、编码容量大，能有效地适应本文提出的基于移位变换的句子层自然语言信息隐藏方法，也可以应用于其他中文文本信息隐藏方法中.

参考文献(References)

[1] Topkara M, Topkara U, Atallah M J. Words are not enough: Sentence level natural language watermarking [C]. Proc of ACM Workshop on Content Protection and Security. Santa Barbara: ACM Press, 2006: 37-46.

[2] Murphy B, Vogel C. The syntax of concealment:

Reliable methods for plain text information hiding[C]. Proc of the SPIE Int Conf on Security, Steganography, and Watermarking of Multimedia Contents. San Jose: ACM Press, 2007: 351-362.

[3] Murphy B, Vogel C. Statistically constrained shallow text marking: Techniques, evaluation paradigm, and results[C]. Proc of the SPIE Int Conf on Security, Steganography, and Watermarking of Multimedia Contents. San Jose: ACM Press, 2007: 363-371.

[4] Meral H M, Sankur B, Ozsoy S. Syntactic tools for natural language watermarking[C]. Proc of the SPIE Int Conf on Security, Steganography, and Watermarking of Multimedia Contents. San Jose: ACM Press, 2007: 339-350.

[5] Atallah M J, Raskin V, Crogan M, et al. Natural language watermarking: Design, analysis, and proof-of-concept implementation[C]. Proc of the 4th Int Inf Hiding Workshop. Berlin Heidelberg: Springer Verlag, 2001: 185-199.

[6] Gupta G, Pieprzyk J, Wang H X. An attack-localizing watermarking scheme for natural language documents [C]. Proc of ASIACCS'06. Taipei: ACM Press, 2006: 157-165.

[7] Topkara M, Riccardi G, Hakkani-Tur D, et al. Natural language watermarking: Challenges in building a practical system[C]. Proc of the SPIE Int Conf on Security, Steganography, and Watermarking of Multimedia Contents. San Jose: ACM Press, 2006: 106-177.

[8] 孙星明，殷建平，陈火旺，等. 汉字的数学表达式研究[J]. 计算机研究与发展，2002，39(6)：707-711.
(Sun X M, Yin J P, Chen H W, et al. On mathematical expression of a Chinese character[J]. J of Computer Research and Development, 2002, 39(6): 707-711.)

[9] 李临定. 汉语比较变换语法[M]. 北京：中国社会科学出版社，1988.
(Li L D. Chinese comparison transform grammar[M]. Beijing: Chinese Social Science Press, 1988.)

(上接第 1855 页)

[4] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Trans on Evolutionary Computation, 2002, 6(2): 182-197.

[5] Li X D. A non-dominated sorting particle swarm optimizer for multiobjective optimization[J]. Lecture Notes in Computer Science, 2003, 2723: 37-48.

[6] Laumanns M, Thiele L, Deb K, et al. Combining convergence and diversity in evolutionary multi-objective optimization[J]. Evolutionary Computation, 2002, 10(3): 263-282.

[7] Mostaghim S, Teich J. The role of ε-dominance in multi-objective particle swarm optimization methods[C]. Proc of IEEE Swarm Intelligence Symposium. Canberra, 2003: 1764-1771.

[8] Sierra M R, Coello C A C. Improving PSO-based multi-objective optimization using crowding mutation and ε-dominance[C]. Int Conf on Evolutionary Multi-criterion Optimization. Guanajuato, 2005: 505-519.

[9] Coello C A C, Pulido G T, Lechuga M S. Handling multiple objectives with particle swarm optimization[J]. IEEE Trans on Evolutionary Computation, 2004, 8(3): 256-279.

[10] Coello C A C, Lechuga M. MOPSO: A proposal for multiple objective particle swarm optimization[C]. Proc of IEEE Congress on Evolutionary Computation. Hawaii, 2002: 1051-1056.

[11] Tripathi P K, Bandyopadhyay S. Adaptive multi-objective particle swarm optimization algorithm[C]. Proc of IEEE Congress on Evolutionary Computation. Singapore, 2007: 2281-2288.

[12] Mostaghim S, Teich J. Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO)[C]. Swarm Intelligence Symposium 2003. Indiana, 2003: 26-33.

[13] Mahfouf M, Chen M Y, Linkens D A. Multi-objective optimal design of alloy steels using adaptive weighted particle swarm optimization[C]. Proc of Parallel Problem Solving from Nature-PPSN Ⅷ. Birmingham, 2004: 762-771.

[14] Kennedy J, Eberhart R. Particle swarm optimization, neural networks[C]. Proc of IEEE Int Conf on Neural Networks. Perth, 1995: 1942-1948.

[15] Shi Y, Eberhart R. A modified particle swarm optimizer[C]. IEEE World Congress on Computation Intelligence. Anchorage, 1998: 69-73.

[16] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach[J]. IEEE Trans on Evolutionary Computation, 1999, 3(4): 257-271.

[17] Corne D W, Jerram N R, Knowles J D, et al. PESA-II: Region-based selection in evolutionary multiobjective optimization[C]. Proc of the Genetic and Evolutionary Computing Conf. San Francisco: Morgan Kaufmann, 2001: 283-290.

[18] Zhang Q F, Zhou A, Jin Y. RM-MEDA: A regularity model-based multiobjective estimation of distribution algorithm[J]. IEEE Trans on Evolutionary Computation, 2008, 12(1): 41-63.