

# An Anti-attack Watermarking Based on Synonym Substitution for Chinese Text

He Lu<sup>1,2</sup>,Lin JianBin<sup>1</sup>,Li TianZhi<sup>1</sup>,Fang DingYi<sup>1</sup>

<sup>1</sup> School of Information Science & Technology  
Northwest University,  
Xi'An, China  
helu1977@yahoo.com.cn,

<sup>2</sup> School of Electrical and Information Engineering  
Xi'An JiaoTong University  
Xi'An, China  
linjianbin116295@163.com

**Abstract**—Current research of natural language steganographic algorithms based on synonymy substitution mostly focused on invisibility, but ignored robustness. However, automatic disambiguation of Chinese word senses(WSD) achieves high accuracy, the adversary could destroyed the watermark easily if he disambiguated the stego-text and did the synonym substitution again. In this paper, against the high accuracy of WSD algorithm, two indicators are proposed, which are Lexical Similarity and Sense Similarity. For reducing the accuracy rate of WSD, we argued that it should choose the word, which is low Lexical Similarity and high Senses Similarity, in the synonym substitution. Therefore, the automatically attack will failure. The experiment showed that the algorithm reduced the accuracy of WSD from 90.4% to 74.5%. The robustness of watermarking has been improved.

**Keywords:** *Information hiding, Natural Language Watermarking, Synonym Substitution, Anti-attack*

## I. INTRODUCTION

Today, so much information on Internet is exchanged in the form of text. For protecting copyright text watermarking technologies are concerned by researchers. In order to embed the invisible watermarking in plain text, Mikhail J. Atallah et.al[1] put forward the concept of Natural Language Watermarking. The so-called Natural Language Watermarking (NL Watermarking for short) are embedding information into semantic structure by using of natural language processing technology to change the lexical, grammar or semantics structure in the text. Meanwhile, preserved the original meaning [2]. The NL Watermarking which used synonym substitution has been paid more attention, since the lexical processing techniques are well developed. T-Lex[3] is a such steganography tool. However, T-Lex doesn't take the context into consideration in the process of synonym substitution. T-Lex has two shortcomings: First, it sometimes replaces words with synonyms that do not agree with correct English usage. Second, T-Lex also substitutes synonyms that do not agree with the genre and the author style of the given text [4]. Therefore, T-Lex concealment algorithm is bad.

In recent years, many researchers have attempted to deal with these problems. Reference [5] makes use of Dependency Parser to get the context of cover text, and decide which synonym to use. Experiment result shows that its concealment and anti-detectability are well. Reference [6] improves the synonym substitution by collocationally-

verified, which requires a specific corpus. By make use of WordNet synonymy corpus and Internet to statistics collocation words, reference [7] put forward a synonym substitution algorithm. The algorithm is effective in style checking.

These algorithms mainly consider the concealment and anti-detection of watermark but concern the robustness little. Equimark [8] quantifies the ambiguity of the embedded information on the original article. The ambiguity in the original text before embedding is less and Equimark select synonyms which have maximum ambiguity to replace the old ones. Because of the imperfect of Word Sense Disambiguation (WSD) techniques, an attacker can not destroy the watermark by automatic attack. However, there are two disadvantages in Equimark: First, the Equimark is based on an imperfect NLP disambiguation technique. In fact, the WSD for English tools are only achieving an accuracy rate of 67.3% to 80% [9]. However, WSD for Chinese tools are achieving an accuracy rate of 90% [10]. Second, cover text requires having been tagged with the word sense and part of speech by hand.

It is a fact that Chinese WSD has been achieved a high accuracy rate. To this point, we put forward an anti-attack watermarking algorithm based on the Chinese synonym substitution. At the same time, we take the concealment into consideration. The structure of paper as follows: we discuss an adversary model in section II. The theory and algorithm of the watermarking embedding and extraction algorithms are proposed in Section III. Experimental results and analysis are in Section IV. Finally is the conclusion.

## II. THE MODEL ADVERSARY

We propose an adversary model based on synonym substitution:

1. Except the secret key, the adversary knows the watermark embedding and extraction algorithm and corpus.

2. The adversary can access the same collocation words algorithm, WSD tools, words dictionary and so on, which were used in embedding algorithm. He can use the same tools to make synonyms substitution on the stego-text. However, the article should not be undermined the original meanings.

We can draw the conclusion from the model above that the adversary can do synonym substitution on the stego-text but not undermine the original intend by the same technology proposed in [3][5][6][7].

From the above analysis of adversary model and the high accuracy rate of Chinese WSD, we argue that in the process of embedding, it can reduce the accuracy rate of WSD tools to stego-text significantly, so that while attacking, synonym substitution may not be correct. When many errors or ambiguity occur during the synonyms substitution, the original meaning of cover-text is crabbed.

### III. EMBEDDING AND EXTRACTION ALGORITHM

#### A. Principle of Algorithm

In fact, the information we possessing and the information adversary grasps is asymmetrical. The cover text has not been processed and the meaning of words is clear. So, the WSD accuracy rate is high. The approximately meaning preserving changes that we make are in the direction of more ambiguity, and WSD is harder for the adversary than it is for us.

At present, the algorithm of Chinese WSD is mainly corpus-based statistical disambiguation method. Corpus-based statistical method selects the word sense which has the greatest weight of probability as the best results by calculating the given word's weight of probability in the context. These methods, such as Maximum Entropy, Bayesian Classification, Hidden Markov model, are all statistical methods. The accuracy rate of WSD depends on the words order and transition probability. After substituted, if there isn't much difference between all senses of the new word and there is a certain semantic difference between the new and the old one, the probability of each sense is very close, and the accuracy rate of statistical disambiguation would decline. We measure the senses difference of the words using lexical similarity and the similar degree of senses using senses similarity.

**Definition 1** Lexical similarity: it measures the ability of two synonyms are interchanged in different context, but aren't changing the syntactic and semantic structure.

Methods of calculating lexical similarity include the literal meaning similarity algorithm, similarity computing by morpheme [11] and similarity computing by semantic taxonomy tree [12] et. al. Among them, the algorithm of calculate lexical similarity based on HowNet is one of the current better methods [12]:

For the two Chinese words  $w_1$  and  $w_2$ , assuming  $w_1$  has  $n$  senses:  $s_{11}, s_{12}, \dots, s_{1n}$  and  $w_2$  has  $m$  senses:  $s_{21}, s_{22}, \dots, s_{2m}$ . We make a provision that the lexical similarity of  $w_1$  and  $w_2$  is decided by the maximum similarity of all senses [12]. Then the lexical similarity of  $w_1$  and  $w_2$ :

$$Sim(w_1, w_2) = \max_{\substack{i=1 \dots n \\ j=1 \dots m}} sim(s_{1i}, s_{2j}) \quad (1)$$

In the *HowNet*[13], the senses are described by sememes. According to hyponymy, all the sememes formed as a layered structure tree. The similarity is measured by calculating the semantic distance. Assuming that the path distance of two sememes in the layered structure denoted with  $d$ , we can get the semantic distance of the two senses:

$$sim(p_1, p_2) = \frac{a}{d + a} \quad (2)$$

$p_1$  and  $p_2$  represent two different senses,  $d$  represent the path distance of  $p_1$  and  $p_2$  in the sememe layered structure, its value is a positive integer. The  $a$  is a variable and  $a = 1.6$  [12].

**Definition 2** Senses similarity: the similarity of all senses of one word.

Assuming that  $w_o$  is a word in cover text,  $s_o$  denotes the senses of  $w_o$  in a context of cover text.  $W$  denotes a synonym set of  $w_o$  when  $s_o$  is the current senses of  $w_o$ .  $w_c$  is a word of the synset  $W$ , that is to say,  $w_c \subseteq W$ . Senses similarity is defined as the expected value of the accumulation of similarity between  $w_c$ 's every senses and  $s_o$ . The number of  $w_c$ 's senses denoted with  $|w_c|$ . Then, the formula to calculate senses similarity of  $w_c$  can be defined as follows:

$$E(w_c, s_o) = \frac{\sum_{s_i \in S(w_c)} sim(s_o, s_i)}{|w_c|} \quad (3)$$

$s_i$  denotes a sense of  $w_c$ .  $sim(s_o, s_i)$  denotes the similarity of the two senses  $s_o$  and  $s_i$  of  $w_c$ , and its value is inversely proportional to the calculation the relative degree of  $s_o$  and  $s_i$ .

#### B. Encoding algorithm

According to the theory discussed above, we design an embedding algorithm. To describe the algorithm clearly, a definitions should be presented first.

**Definition 3** substitution queue: all substitutable words in cover text which sorted by secret key composed a queue.

This algorithm takes three inputs: a secret key, a message  $M$ , and cover text. Steps of embedding process are as follows:

- 1) Using Chinese WSD tools to tag cover-text. Then, construct substitution queue by the secret key.
- 2)  $C = 1$ ;
- 3) For each word  $w_i$  in substitution queue:
  1.  $bit_c = M[C]$ .
  2. Look up synset  $W$  of  $w_i$  in the dictionary, according to the results of disambiguation.
  3. For each word  $w_{ij}$  in  $W$ :
    - a. Calculate senses similarity  $E$  of  $w_{ij}$  and lexical similarity  $sim$  of  $w_{ij}$  and  $w_i$ .
    - b. According to  $E$ , rank  $W$  in descending order, remove synonyms in  $W$  which  $sim$  larger than threshold  $a$  or  $sim$  small than threshold  $b$ .
    - c. Encode  $w_{ij}$  "0" or "1" by the key. Neighborhood synonyms should not assign the same code.
    - d. Increment  $C$ .
4. If  $C = |M|$ , end the cycle.

#### C. Extraction Algorithm

Steps of watermark extraction process are as follows:

- 1) Using Chinese WSD tools to tag the stego-text. Then, construct substitution queues No.1 and No.2 by the secret key.

- 2)  $C=1$ ;
- 3) For each words  $w_{1i}$  and  $w_{2i}$  in substitution queues No.1 and No.2:
  1. Look up synset  $W_1$  of  $w_{1i}$  in the dictionary, according to the results of disambiguation.
  2. For each word  $w_{1ij}$  in  $W_1$ :
    - a. Calculate senses similarity  $E$  of  $w_{1ij}$  and lexical similarity  $sim$  of  $w_{1ij}$  and  $w_{1i}$ . According to  $E$ , rank  $W_1$  in descending order, remove synonyms in  $W_1$  which  $sim$  larger than threshold  $a$  or  $sim$  small than threshold  $b$ .
    - b. Encode  $w_{1ij}$  "0" or "1" by the key. Neighborhood synonyms should not assign the same code.
    - c. extract the watermark. Look for  $w_{1ij}$  as the same word as  $w_{2i}$  in  $W_1$ . If  $w_{1ij}$  encode "1",  $M[C]=1$ , If  $w_{1ij}$  encode "0",  $M[C]=0$ .
    - d. Increment  $C$ .
  3. If  $C = |M|$ , end the cycle.

#### IV. EXPERIMENTAL RESULTS AND ANALYSIS

We design experiment according to the algorithm proposed in section III. We have selected 25 test texts which lengths are range from 1000 to 2000 words. We have used HIT Chinese WSD tool [14] to tag texts and generated a substitution queue. The synset were taken from the "HIT-IR Tongyici Cilin (Extended) dictionary" [14]. The lexical similarity threshold  $a$  and  $b$  are 80% and 50% in the watermarking algorithm.

中方对此高度关注，要求俄方抓紧搜救失踪船员，认真负责地对事故原因从速进行彻底、详细调查，并尽快将结果通报中方。

Figure 1. Part of cover-text

中方对此高度关切，要求俄方抓紧搜救失踪船员，认真负责地对事故原委从速进行彻底、详细调查，并尽快将结果通知中方。

Figure 2. Part of stego-text

中方对此高度关怀，要求俄方抓紧搜救失踪船员，认真负责地对事故源流从速进行彻底、详细调查，并尽快将结果通报中方。

Figure 3. Part of text after attack

The original cover-text as showed in Figure 1. The text after embedding watermark is showed in Figure 2. And the text after attack is showed in Figure 3.

According to our statistical data, the average accuracy rate of disambiguation before embedded watermark is about 90.4% and after watermark embedding it is about 74.5%.

We can draw conclusions: first, after embedding the

accuracy rate of WSD reduced from 90.4% to 74.5%. If the adversary substitutes synonyms again, it will surely result in that the new words don't coherent in context. Therefore, the algorithm is of good robustness. Second, Comparing Figure 1 with Figure 2, we can see that the new words after first substitution don't change the original meaning of text. In fact, the algorithms consider the substituted words' semantic context in the cover text in the process of synonyms substitution so that it excludes the incorrect synonym replacement. The algorithm is of good concealment.

#### V. CONCLUSIONS

In this paper we propose an anti-attack watermarking algorithm based on Chinese synonyms substitution. By substituting the synonyms which are low in lexical similarity but high in senses similarity, the algorithm can reduce the accuracy rate of the WSD tools and enhance the anti-attack ability of text watermarking. The result of the experiments indicates that this algorithm is robust.

#### ACKNOWLEDGMENT

This work was supported by national university of innovative experimental projects (200722).

#### REFERENCES

- [1] Atallah M, McDonough C, Nirenburg S, Raskin V. Natural language processing for information assurance and security: an overview and implementations. New Security Paradigms Workshop. Ireland, 2001: 51-65.
- [2] M. Atallah, V. Raskin, C. Hempelmann, M. Karahan, R. Sion, and K. Triezenberg, Natural language watermarking and tamperproofing, Lecture Notes in Computer Science and Proceedings of the 5th International Information Hiding Workshop, 2002.
- [3] Keith Winstein, Tyrannosaurus Lex.Website, <http://alumni.imsa.edu/~keithw/tlex>. accessed 2008.10
- [4] C. Taskiran, U. Topkara, M. Topkara, and E. J. Delp, "Attacks on Lexical Natural Language Steganography Systems," SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents, San Jose, CA, 2006.
- [5] Gan Can, Sun Xingming, Liu Yuling, Xiang Lingyun. An improved steganographic algorithm based on synonymy substitution for Chinese text, 2007,37(1S): 137-140.
- [6] Igor A. Bolshakov. A method of linguistic steganography based on collocationally-verified synonymy. In Jessica J. Fridrich, editor, Information Hiding: 6th International Workshop, 2004, Volume 3200,180-191.
- [7] I.A.Bolshakov, A. Gelbukh. Synonymous Paraphrasing Using WordNet and Internet. Springer Berlin, Heidelberg,2004, Volume 3136,312-323
- [8] U. Topkara, M. Topkara, M. J. Atallah. The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions. Proceedings of ACM Multimedia and Security Workshop (MMSEC'06), Geneva, Switzerland, 2006.9, 26-27.
- [9] JIN Peng, WU Yun-fang, YU Shi-wen. Survey of Word Sense Annotated Corpus Construction. Journal of Chinese information processing. 2008,22 (03) : 16-23.
- [10] Harbin Institute of Technology Information Retrieval Lab. Website, [http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE\\_user\\_op=view\\_page&PAGE\\_id=146&MMN\\_position=51:48](http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=146&MMN_position=51:48)

- [11] ZHU Yi-hua, HOU Han-qing, SHA Yin-ting. A Comparison of Two Algorithms for Computer Recognition of Chinese Synonyms. Journal of Library Science in China. 2002, 28(140): 82-85.
- [12] Qun LIU, Sujian LI. Word Similarity Computing Based on How-net. The Third Chinese Lexical Semantics Workshop. TaiBei, 2002.
- [13] Dong Zhengdong, Dong Qiang. <http://www.keenage.com/>
- [14] IRLab . <http://ir.hit.edu.cn>