ІНФОРМАТИКА, ОБЧИСЛЮВАЛЬНА ТЕХНІКА ТА АВТОМАТИЗАЦІЯ

УДК 003.26

Тарасенко Я.В.
Черкаський державний технологічний університет

МОДЕЛЮВАННЯ СИНТАКСИЧНИХ СТРУКТУР ДЛЯ АТАКИ СЕМАНТИЧНИМ СТИСНЕННЯМ НА ЛІНГВІСТИЧНУ СТЕГОСИСТЕМУ

Формалізується задача отримання вхідних даних для дискурсного аналізу в процесі синтаксичного дослідження для атаки семантичним стисненням на лінгвістичну стегосистему. Конкретизується метод моделювання стеганографічних об'єктів для потреб та задач комп'ютерного лінгвістичного стегоаналізу. Подальшого розвитку набуває підхід синтаксичного аналізу для інформаційно-пошукової системи. Робиться висновок про можливе однотипне застосування моделювання для проведення атаки на будь-яку лінгвістичну стегосистему, що заснована як на морфологічних чи семантичних, так і на методах довільних інтервалів. Доводиться ефективність подібного моделювання у програмному комплексі проведення атаки на лінгвістичну стегосистему за рахунок інтеграції у середовище дискурсного аналізу.

Ключові слова: *дискурсний аналіз, атака семантичним стисненням, лінгвістична стегосистема, моделювання стегооб'єктів, комп'ютерний стегоаналіз, синтаксичний аналіз.*

Постановка проблеми. Зі зростанням кількості текстової інформації, що передається мережею Інтернет, зростає імовірність несанкціонованого обміну даними та витоку секретної інформації, інакше кажучи, набуває розвитку та поширення текстова стеганографія. Той факт, що найбільш популярний сьогодні напрям вбудовування стегоповідомлення у зображення зі значною надлишковістю [1] сприяє розвитку методів стегоаналізу та проведення атак на стегосистему саме у цьому напрямі. Звідси слідує, що текстовій стеганографії приділено значно менше уваги, тому імовірність вдалого передання стегоповідомлення, використовуючи текст контейнером значно збільшується.

Аналіз останніх досліджень і публікацій. В.Г. Грибунін в [2] описує математичну модель стегосистеми взагалі, яку можна застосовувати до конкретних часткових випадків. Що стосується стегоаналізу саме текстової інформації, лінгвістичної стеганографії, то в цьому разі з методами, заснованими на дослідженні імовірнісно-статистичних характеристик [3; 4] чи методами, що базуються на частотному дослідженні [5], не менш важливу групу теж складають методи, пов'язані з математичним моделюванням [6]. П.І. Аношин одним із двох підходів до вирішення проблеми синтаксичного аналізу виділяє імовірнісно-статистичний [7]. Це доводить той факт, що будь-який лінгвістичний аналіз тексту, зокрема синтаксичний (за основу моделі синтаксичної структури можна взяти дерево прийняття рішень, описане в [8]) є сам по собі атакою на текстову стегосистему.

Моделювання досить часто зустрічається в лінгвістиці для формального опису певного лінгвістичного процесу чи явища [9]. Водночас формалізації потребує процес отримання інформації під час синтаксичного стеганографічного дослідження.

Є.В. Разінков у дисертації на тему «Математичне моделювання стеганографічних об'єктів і методи вирахування оптимальних параметрів стегосистем» [10] стверджує, що чим більш чітка модель стегоконтейнерів є у фахівця стеганографа, тим більш стійку стегосистему він може побудувати, проте це твердження правдиве і коли йдеться про зворотну ситуацію, де ефективність стегоатаки пропорційно залежить від якості моделі стегосистеми, наявної у стегоаналітика [10]. У разі автоматизованого стегоаналізу це стосується ефективності побудови

моделі стегосистеми програмою. Тобто система стегоаналізу повинна будувати декілька імовірних стегосистем та порівнювати результати з досліджуваним текстом. Так, можливо визначити метод приховування повідомлення, тобто існуватиме шанс його розшифрування. В іншому випадку повідомлення буде видалено, що змусить зловмисника повторити передання, використовуючи інший алгоритм. Актуальність дослідження зумовлена необхідністю розроблення підходів для проведення атаки проти вбудованого повідомлення [2] у текст природної мови за допомогою лінгвістичних засобів та реалізації цих підходів і методів у програмному комплексі проведення атаки на лінгвістичну стегосистему [11]. Для більш ефективного опису методу необхідним є формалізація певних процесів, що є основою для проведення атаки семантичним стисненням.

Подальшого розвитку набуває метод синтаксичного аналізу [8], зокрема адаптується для врахування можливості застосування синтаксичних засобів стеганографії.

Постановка завдання. Метою статті є формалізація задачі отримання вхідних даних для дискурсного аналізу в процесі синтаксичного дослідження для проведення атаки семантичним стисненням на лінгвістичну стегосистему. Крім того, на прикладі моделювання синтаксичних структур необхідно дослідити особливості та переваги такого підходу для атаки на будь-яку іншу лінгвістичну стегосистему в програмному комплексі, що забезпечує протидію широкому спектру загроз, спричинених методами лінгвістичної стеганографії.

Для досягнення мети перед роботою були поставлені такі завдання:

1) визначення моделі можливих синтаксичних структур лінгвістичної стегосистеми;

2) визначення дій, у разі повного та неповного співпадіння змодельованої стегосистеми та досліджуваної структури;

3) дослідження шляху остаточного прийняття рішення про наявність прихованого повідомлення у конструкції.

Виклад основного матеріалу дослідження. Одне з можливих результативних застосувань моделювання стосується атаки на лінгвістичну стегосистему, в якій використано синтаксичні методи стеганографії. Така атака базується на імовірнісно-статистичному дослідженні синтаксичної структури, доведення ефективності якого для задоволення потреби синтаксичного аналізу в межах задач стеганографії описано в [12]. Практичне використання та застосування цього підходу описано в [8]. Крім того, синтаксична синонімія, що визначається на етапі синтаксичного аналізу, також основана на моделюванні імовірних синтаксичних конструкцій одного і того ж речення. Основною особливістю такого підходу є інтеграція у середовище дискурсного аналізу за рахунок побудови дерева відповідності та спрощення процесу прийняття рішень щодо дій, направлених на видалення повідомлення програмними засобами, які реалізують автоматизовану атаку на лінгвістичну стегосистему [11].

Можна взяти за основу найпростіший випадок використання дерева прийняття рішень (рис. 1) для синтаксичного аналізу, описаного у [8].

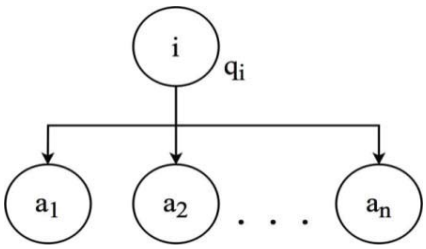

Рис. 1. Найпростіше дерево прийняття рішень

де q_i – деяке питання, на яке відповідає вузол і, a_n можливі відповіді. Адже, що стосується побудови імовірнісного дерева відповідностей під час визначення синтаксичної структури, слід будувати також декілька імовірних дерев, наприклад, у [13], кожне з яких відповідає можливості застосування тих чи інших синтаксичних прийомів стеганографії. Інакше кажучи, проводити моделювання імовірної синтаксичної атаки і порівняння досліджуваного речення з побудованими моделями. Також необхідно врахувати, що генерацію моделі стегосистеми в загальному випадку можна описати формулою (1), що згадується у [2].

$$F : I^* \times K^* \times B^* \to W^*, \qquad (1)$$

де *W*, K*, I*, B** – множини можливих цифрових водяних знаків, ключів, контейнерів та стегоповідомлень відповідно. Тоді побудова дерева відповідностей з урахуванням властивостей стеганографії можна зобразити використовуючи декілька послідовних дерев відповідності, кожне з яких моделює можливу синтаксичну структуру під час використання тих чи інших методів стеганографії. Разом дерева відповідності (рис. 2) утворюють систему, що допомагає визначити імовірність використання синтаксичних методів стеганографії.

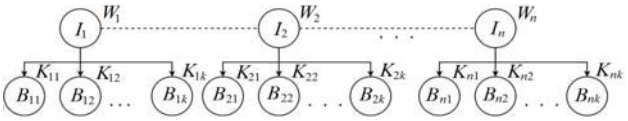

Рис. 2. Дерева відповідностей при моделюванні можливих стегосистем

W_n – множина n можливих стегосистем у досліджуваному реченні; n – кількість можливих синтаксичних структур; k – кількість можливих гілок в кожному з n дерев; B_{nk} – стегоповідомлення, приховане у відповідному члені речення; K_{nk} – імовірний ключ для приховування повідомлення B_{nk}; – I_n контейнер, відповідний вузлу у реченні. У такому разі моделююча стегосистему функція буде мати вигляд (2).

$$F_n : I_n \times K_{nk} \times B_{nk} \to W_n \qquad (2)$$

Отже, кожне досліджуване речення, для якого будується дерево відповідності порівнюється з моделлю можливої стегосистеми, що описується функцією і у разі, коли порядок слів у реченні, його структура повністю відповідає змодельованій стегосистемі, побудованій із використанням структурних елементів цього речення, робиться висновок про можливість приховування повідомлення у ньому. Коли синтаксична структура досліджуваного речення співпадає зі змодельованим лише частково, виникає необхідність виділення найменш імовірних місць приховування стегоповідомлення. Цього можна досягти під час дослідження від зворотного, тобто порівняти усі змодельовані стегосистеми між собою і, знайшовши спільне, визначити вірогідну модель стегосистеми (рис. 3).

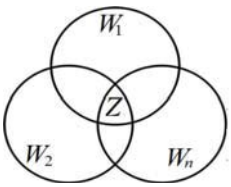

Рис. 3. Визначення вірогідної моделі стегосистеми

Остаточне рішення приймається за допомогою використання кругів Ейлера, де W_1, W_2,…,W_n – моделі стегосистеми, за кожну з яких відповідає свій круг. Моделі відрізняються застосованим синтаксичним методом стеганографії та місцем, де можливо приховати повідомлення. Z – область збігу змодельованих стегосистем. У такому разі можна зробити висновок, що спільні незмінні елементи для усіх стегомоделей є тими елементами, які можуть бути модифіковані засобами стеганографії з найменшою імовірністю, оскільки використання іншого методу чи моделі стегосистеми означатиме зміну засобів приховування повідомлення, тому зміну лексичних одиниць і (або) синтаксичної структури окремих речень чи семантичної структури усього тексту.

Z визначається за формулою (3) перетином n змодельованих множин, кожна з яких є можливою синтаксичною стегосистемою.

$$Z = W_1 \cap W_2 \cap \ldots \cap W_n \qquad (3)$$

Виявлення не лише факту наявності стегоповідомлення, а і методу та точного місця його приховування дає змогу не тільки його видаляти, але і змінювати чи розшифровувати, знайшовши ключ. Моделювання не може стовідсотково гарантувати виявлення повідомлення. Наприклад, у роботі [6] зазначається точність стегоаналізу, що дорівнює 92%. Таким чином, можна стверджувати про наявність помилок як першого, так і другого роду. Проте в комплексі з іншими засобами стегоаналізу, що підвищують цю точність, ефективність проведення атаки зростає.

Очевидно, процес моделювання може негативно відобразитись на комп'ютерній системі, яка реалізує автоматизований стегоаналіз у вигляді збільшення часу оброблення інформації. Проте ефективність будь-якої атаки компенсує ці негативні прояви та відкриває широкий спектр можливих застосувань моделювання стегооб'єктів для атаки на будь-яку лінгвістичну стегосистему.

Висновки. У роботі формалізовано задачу отримання вхідних даних для дискурсного аналізу в процесі синтаксичного дослідження для проведення атаки семантичним стисненням на лінгвістичну стегосистему.

Під час роботи визначено модель можливих синтаксичних структур лінгвістичної стегосистеми, що є деревами відповідностей, змодельованими на основі потенційних структур досліджуваного речення, у яких можливе приховування стегоповідомлення тим чи іншим методом синтаксичної стеганографії.

У разі повного співпадіння досліджуваного та змодельованого речення робиться висновок про високу імовірність використання параметрів, що застосовувались під час моделювання цього речення, у разі часткового збігу, точність стегоаналізу підвищується за рахунок застосування кругів Ейлера, на основі чого приймається остаточне рішення про наявність стегоповідомлення у реченні.

Крім того, на прикладі моделювання синтаксичних структур досліджено можливість використання такого підходу за умов атаки на будь-яку іншу лінгвістичну стегосистему для застосування

у програмному комплексі, що забезпечує протидію широкому спектру загроз, спричинених методами лінгвістичної стеганографії. Виявлено, що запропонований у статті підхід найкраще інтегрується у середовище дискурсного аналізу за рахунок побудови дерева відповідності та спрощує процес прийняття рішень щодо дій, направлених на видалення повідомлення комп'ютерними системами, які реалізують автоматизовану атаку на лінгвістичну стегосистему.

Отримані результати дають можливість практично застосовувати метод моделювання стеганографічних об'єктів та імовірнісний синтаксичний аналізатор для потреб комп'ютерного лінгвістичного стегоаналізу тексту, модифікованого засобами синтаксичної стеганографії, а також підтверджують можливе ефективне застосування для проведення атаки на стегосистему, основану на морфологічних, семантичних методах чи методах довільних інтервалів та протидії їм у програмному комплексі проведення атаки на лінгвістичну стегосистему шляхом видалення будь-якого стегоповідомлення, наявного у тексті.

Список літератури:

1. Ахмаметьєва Г.В., Мурова В.В. Удосконалення стегоаналітичного методу виявлення вкладень додаткової інформації в цифрових зображеннях, заснованого на аналізі послідовних тріад колірних триплетів. Інформатика та математичні методи в моделюванні. 2017. Том 7. № 3. С. 187–194.

2. Грибунин В.Г., Оков И.Н., Туринцев И.В. Цифровая стеганография. Москва: СОЛОН-ПРЕСС, 2009. 263 с.

3. A Statistical Algorithm for Linguistic Steganography Detection Based on Distribution of Words / Z. Chen, L. Huang, Z. Yu [at al.] // Availability, Reliability and Security: proceedings of III international conference (Barcelona, Spain, 4–7 March 2008). Berlin, 2008. P. 558–563.

4. Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words / Z. Chen, L. Huang, Zh. Yu [et al.] // Information Hiding: revised selected papers of X international workshop, IH 2008 (Santa Barbara, USA, 19–21 May 2008). Berlin, 2009. Vol. 5284. P. 224–235.

5. Chen Z., Huang L., Yang W. Detection of substitution-based linguistic steganography by relative frequency analysis. Digital investigation. 2011. № 8 (1). P. 68–77.

6. Linguistic Steganography Detection Based on Perplexity / P. Meng, L. Huang, Z. Chen [et al.] // Proceedings of International Conference on MultiMedia and Information Technology (Three Gorges, China, 30–31 December 2008). Berlin, 2008. P. 217–220.

7. Аношин П.И. Автоматический анализ текстов. Синтаксический и семантический анализ. Евразийский научный журнал. 2017. № 6. С. 211–213.

8. Андреев А.М., Березкин Д.В., Брик А.В., Кантонистов Ю.А. Вероятностный синтаксический анализатор для информационно-поисковой системы. Компьютерная хроника. 1999. № 1. С. 37–85.

9. Поветкина Ю.В. Моделирование как метод лингвистического исследования. Филологические науки. Вопросы теории и практики. 2012. № 6 (17). С. 132–136.

10. Разинков Е.В. Математическое моделирование стеганографических объектов и методы вычисления оптимальних параметров стегосистем: автореф. дис. … канд. фіз.-мат. наук : 05.13.18. Казань, 2012. 17 с.

11. Тарасенко Я.В. Програмний комплекс проведення атаки на лінгвістичну стегосистему. Безпека інформації. 2018. № 24 (1). С. 56–61.

12. Федотова-Півень І.М., Тарасенко Я.В. Шляхи задоволення потреб сучасної кібербезпеки в рамках протидії методам комп'ютерної лінгвістичної стеганографії. Безпека інформації. 2017. № 23(3). С. 190–196.

13. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural language Toolkit. Sebastopol: O'Reilly Media, 2009. 504 p.

МОДЕЛИРОВАНИЕ СИНТАКСИЧЕСКИХ СТРУКТУР ДЛЯ АТАКИ СЕМАНТИЧЕСКИМ СЖАТИЕМ НА ЛИНГВИСТИЧЕСКУЮ СТЕГОСИСТЕМУ

Формализуется задача получения входных данных для дискурсного анализа в процессе синтаксического исследования для атаки семантическим сжатием на лингвистическую стегосистему и конкретизируется метод моделирования стеганографических объектов для нужд и задач компьютерного лингвистического стегоанализа. Дальнейшее развитие получает подход синтаксического анализа для информационно-поисковой системы. Делается вывод о возможном однотипном применении моделирования для проведения атаки на любую лингвистическую стегосистему, основанную как на морфологических или семантических так и на методах произвольных интервалов. Доказывается эффективность подобного моделирования в программном комплексе проведения атаки на лингвистическую стегосистему за счет интеграции в среду дискурсного анализа.

Ключевые слова: *дискурсный анализ, атака семантическим сжатием, лингвистическая стегосистема, моделирование стегообъектов, компьютерный стегоанализ, синтаксический анализ.*

SYNTACTIC STRUCTURE SIMULATION FOR SEMANTIC COMPRESSION-BASED ATTACK ON THE LINGUISTIC STEGOSYSTEM

The work is associated with the formalization of the input data obtaining task for discursive analysis in the process of syntactic research for carrying out an attack by semantic compression on a linguistic stegosystem and specification of the steganographic objects simulation method for the needs and tasks of computer linguistic steganalysis. The further development of the syntactic analysis approach for the information retrieval system is carried out. It is made a conclusion that a simulation may be used for the same kind of attack on any linguistic stegosystem that is based either on morphological or semantic methods or on random interval methods. The effectiveness of such an application in the program complex for attacking the linguistic stegosystem due to integration into the environment of discourse analysis is proved.

Key words: *discursive analysis, attack by semantic compression, linguistic stegosystem, modeling of stegoobjects, computer steganalysis, syntactic analysis.*