УДК 003.26
DOI: 10.30748/nitps.2018.32.10

Я.В. Тарасенко, О.Б. Півень, І.М. Федотова-Півень

Черкаський державний технологічний університет, Черкаси

МЕТОД СЕМАНТИЧНОГО СТИСНЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ ДЛЯ ПРОТИДІЇ КОМП'ЮТЕРНІЙ ЛІНГВІСТИЧНІЙ СТЕГАНОГРАФІЇ

В статті наводиться розробка методу семантичного стиснення текстової інформації для протидії комп'ютерній лінгвістичній стеганографії, що складається з п'яти послідовних етапів (автоматизованого лінгвістичного аналізу, оцінки осмисленості тексту та виділення основної думки, стиснення, модифікації та формування остаточного тексту після внесення змін) на основі дискурсного аналізу, що визначається адаптацією методів реферування тексту для врахування можливого використання засобів стеганографії та проведення атаки на лінгвістичну стегосистему шляхом семантичного стиснення як осмислених, так і штучно згенерованих текстів на основі комплексного підходу та врахування широкого спектру засобів стеганографії.

Ключові слова: *комп'ютерна лінгвістична стеганографія, семантичне стиснення тексту, текстовий стегоаналіз, автоматизований стегоаналіз, метод стиснення тексту, визначення осмисленості тексту, атака на лінгвістичну стегосистему.*

Вступ

Постановка проблеми. У сучасному інформаційному суспільстві, за умови зростання ризику терористичних атак, промислового шпигунства та стану, в якому знаходиться країна на поточному етапі існування постає гостра необхідність в ефективних методах та програмних засобах і системах, що направлені на виявлення та протидію прихованому витоку інформації чи передачі секретних даних за допомогою методів стеганографії. На сьогоднішній день існує умовний поділ стеганографії на класичну, комп'ютерну та цифрову [1]. Комп'ютерна є найбільш перспективним напрямком, що спричинено міграцією файлів у мережі, а отже виникає необхідність розробок комп'ютерних засобів протидії існуючим та перспективним методам приховування повідомлення на основі обробки текстової інформації, а звідси виходить, що слід розглядати лінгвістичні методи та засоби модифікації текстового повідомлення. Якщо простежити історичний розвиток лінгвістичної стеганографії, а відповідно і засобів протидії їй, можна стверджувати, що перспективи їх подальшого розвитку зосереджені саме в напрямку розробки комп'ютерних систем для розуміння смислу та підтексту за допомогою автоматизованого дискурсного аналізу. В той же час, будь-яка атака на стегосистему будується на основі підходів, основною проблемою яких є використання їх у комп'ютерних системах, а атака проводиться на основі результатів стегоаналізу. Особливості стегоаналізу саме текстової інформації зумовлені особливостями природньої мови. Основна задача протидії текстовій стеганографії полягає у проведенні первинного аналізу на найбільш явні сліди модифікації тексту. Для забезпечення виконання задачі проведення атаки на лінгвістичну стегосистему, направлену на видалення прихованого повідомлення, необхідним є створення методу семантичного стиснення текстової інформації, що базується на дослідженні дискурсу тексту та, враховуючи початкову семантику та морфологічно-синтаксичну структуру, проводить його стиснення з втратами, що відзначається розумінням смислу тексту та збереженням його початкової семантики.

Аналіз останніх досліджень і публікацій. Майже весь ринок вітчизняних комп'ютерних систем стегоаналізу займають системи стегоаналізу зображень, деяка частина ринку присвячена стегоаналізу аудіо чи відео, проте існує надзвичайно мало розробок в області комп'ютерного стегоаналізу тексту. Звісно, існує багато закордонних методів та алгоритмів виявлення прихованого повідомлення в тексті, що протидіють тим чи іншим методам текстової стеганографії [2–5], проте, окрім важкої доступності, вони володіють багатьма недоліками, як, наприклад, відсутність можливості врахування семантики, нездатність розпізнати неосмисленість тексту, а звідси витікає неможливість і проведення дій з неосмисленим текстом, що необхідно для проведення атаки на стегосистему, основану на передачі штучно згенерованих текстів. Звісно існують різноманітні поштові сканери, направлені на аналіз повідомлень, проте смислове навантаження та мету його написання вони не в змозі виділити. Крім того, атаки на лексичні стеганографічні системи природньої мови недостатньо вивчені [6]. В той же час, існуючі програмні засоби лінгвістичного аналізу [7], дискурсного аналізу [8–9] та стиснення тексту [10–11] хоч і можна застосовувати для проведення атаки на лінгвістичну стегосистему, проте у зв'язку зі сво-

єю неспеціалізованістю, лише як частину дослідження, що проводить людина, а тому, незважаючи на наявність програм, що реалізують той чи інший метод, спостерігається брак як комерційних продуктів, так і дослідницьких розробок, які б повністю автоматизували процес стегоаналізу тексту. В існуючих розробках відсутній комплексний підхід та наявна направленість на конкретні часткові випадки. Імовірно, недостатність програмних розробок обумовлена тим, що онтологічний підхід [12] стеганографії лише розвивається, тому і методи аналізу та протидії недостатньо розвинені. Але і можливість автоматизованого програмного видалення семантично прихованого стегоповідомлення наразі досліджено не в повній мірі. Цей факт зумовлює відсутність ефективної протидії імовірній зовнішній загрозі, що базується на використанні онтологічного підходу в лінгвістичній стеганографії та відкриває шляхи до майже безперешкодної прихованої передачі даних у тексті, а отже існує необхідність в розробці методів комплексного виявлення факту передачі повідомлення та проведенні відповідних дій по його нейтралізації. Звідси витікають невирішені питання, що зумовлюють актуальність дослідження: необхідність підвищення рівня інформаційної безпеки держави, потреба адаптації процесу автоматизованого лінгвістичного стегоаналізу для проведення атаки стисненням на стегосистему, необхідність вилучення з процесу визначення осмисленості тексту людського фактору, направленість ринку автоматизованих систем стегоаналізу на дослідження зображень, аудіо чи відеофайлів та недостатність розробок в області текстового стегоаналізу, відсутність можливості врахування семантики тексту та нездатність розпізнати неосмислений текст існуючими комп'ютерними системами, недостатня вивченість атак на лексичні стегосистеми, відсутність комплексного підходу в існуючих розробках проведення атаки на лінгвістичну стегосистему та направленість на окремі часткові випадки, неможливість автоматизованого програмного видалення стегоповідомлення, що приховано семантичними засобами, відсутність ефективних засобів протидії онтологічному підходу у лінгвістичній стеганографії. В роботі [13] набула розвитку побудова ущербних текстів на основі ущербних кодів, а нанесення шкоди тексту відбувається на основі видалення символів та обрахування символьної надлишковості тексту для вирішення задач формування багатоканальної криптографії, проте такий підхід не враховує початкову семантику тексту, крім того, можливі втрати основного смислу тексту, що не підходить для виконання поставлених практичних задач інформаційної безпеки, в які входить сканування та модифікація повідомлень зі збереженням семантики та структури початкового тексту.

Мета статті – розробка методу стиснення текстової інформації для протидії методам лінгвістичної стеганографії шляхом реалізації науково-практичної задачі проведення атаки стисненням на лінгвістичну стегосистему. Для досягнення мети було поставлене наступне завдання: розробити метод, що складається з п'яти етапів (автоматизованого лінгвістичного аналізу з урахуванням методів стеганографії, оцінки осмисленості тексту та виділення основної думки, семантичного стиснення, модифікації та формування остаточного тексту після внесення змін) та забезпечує:

1) комплексний стегоаналіз текстових даних (для контролю над процесом атаки стисненням);

2) проведення атаки на текстову стегосистему шляхом стиснення та модифікації з урахуванням початкової семантики тексту та можливої присутності контейнеру з прихованим повідомленням (для видалення стегоповідомлення, і таким чином протидії методам комп'ютерної лінгвістичної стеганографії);

3) виявлення та стиснення неосмислених штучно згенерованих текстів (для врахування усього спектру можливих стегоконтейнерів).

Виклад основного матеріалу

В представленій роботі було використано: математичні методи, до яких належить теорія імовірності, математична статистика та моделювання, теорія множин, метод вирішення задач за допомогою кругів Ейлера та лінгвістичні методи, як морфологічний (метод логічного множення), синтаксичний (імовірнісно-статистичний метод) та дискурсний аналіз (текстуально-інтертекстуальний підхід), методи реферування та написання переказів. Математичні методи з лінгвістичними поєднані використанням області некласичних логік – інтенсіональної логіки та елементів формальної граматики і семантики. Більш детально причини вибору засобів автоматизованого лінгвістичного аналізу тексту та їх переваги для розробки методу стиснення текстової інформації для лінгвістичної стеганографії описані в роботі [14]. Метод складається з п'яти послідовних етапів.

Етап автоматизованого лінгвістичного аналізу тексту. Задача автоматизованого стиснення текстової інформації передбачає проведення, перш за все, виділення основного смислу, шляхом дискурсного аналізу, який в свою чергу не можливий без даних, отриманих в ході синтаксичного аналізу. На цьому етапі методу набуває подальшого розвитку автоматизований морфологічний та синтаксичний аналіз [7; 15; 18], характерною особливістю чого є модифікація для врахування можливості використання методів лінгвістичної стеганографії при аналізі. Для здійснення синтаксичного аналізу слід дослі-

дити кожне слово тексту морфологічно для отримання можливості протидії методам стеганографії, основаним на заміні синонімів, та усієї множині методів, що впливають на морфологічну структуру, а також синтаксичним методам. Отримані дані про частини мови кожного слова та його форму шляхом пошуку відповідника у словнику словоформ, дають зрозуміти рівень правильності вживання тих чи інших слів у тексті, а це означає, що велика кількість помилок у словах може свідчити про можливу наявність стегоповідомлення, приховування якого основане на помилках при автоматичному перекладі тексту [12]. Дослідження статистичного розподілу цих помилок дозволить зробити висновок про їх можливе походження. Факт випадкової появи помилки, що є не типовою для усього тексту, та відсутність цієї помилки у місці, де вона повинна була з'явитись, доводить штучність внесення її до тексту. Таким чином, розподіл помилок у тексті позначимо через відношення $N=\frac{\varphi}{\alpha}$, де φ – кількість типових помилок, α – кількість можливих помилок того ж типу. У випадку, коли помилки у тексті зумовлені особливостями автоматизованого перекладу, N = 1.

Після морфологічного аналізу проводиться синтаксичний аналіз тексту на основі даних, отриманих в ході морфологічного дослідження. Це досягається в результаті моделювання синтаксичних структур на основі підходу логічного множення, причину вибору якого описано в [14]. Детально підхід до синтаксичного аналізу описано в [15] та не потребує дублювання. Особливістю використання цього підходу в роботі є його адаптація для задач стеганографії. Таким чином, при побудові дерева відповідностей слід аналізувати особливості пунктуації. Дані, отримані при побудові імовірних структур речення, передаються для аналізу на етапі дискурсного дослідження тексту. Для врахування можливого використання синтаксичних методів стеганографії, що не береться до уваги в [15], досліджується статистичний розподіл підозрілих речень, а дані передаються дереву прийняття рішень, що відповідає за стегоаналіз. Таким чином, визначається імовірність використання засобів стеганографії у кожному реченні за формулою 1.

$$P(A)=\frac{|A|}{N}, \qquad (1)$$

де N – загальна кількість різних порядків слів у реченні; A – подія, що характеризує наявність стегоповідомлення у реченні; $|A|$ – незвичний порядок слів, що можливо є прихованим повідомленням.

Для запобігання помилки 1-го роду, слід аналізувати весь текст в цілому з урахуванням кожного речення. В такому разі, імовірність того факту, що текст є модифікованим за допомогою стеганографічних засобів, визначається як відношення суми усіх речень, імовірність модифікації яких становить більше 0,5, до загальної кількості речень у тексті (2).

$$P'(A')=\frac{\sum_{i=1}^{n} P_i(A_n)}{|\Omega'|}, \qquad (2)$$

при $P_i(A_n)\geq 0,5$, де $P'(A')$ – імовірність події, що характеризує наявність стегоповідомлення у тексті; $P_i(A_n)$ – імовірність наявності стегоповідомлення у n реченнях; $|\Omega'|$ – загальна множина речень тексту. Далі імовірність модифікації усього тексту, як і окремих речень, записується та передається для наступної обробки на етапі видалення стегоповідомлення та є орієнтиром, які саме місця у тексті слід стискати ретельніше.

Етап оцінки осмисленості тексту та виділення основної думки. Для проведення оцінки осмисленості тексту, визначення наявності слідів стеганографічної модифікації тексту семантичними методами, підтвердження чи спростування факту штучної генерації тексту для задач стеганографії та для вірного видалення стегоповідомлення на етапі стиснення слід аналізувати текст дискурсно. Необхідним є визначення його семантичної цілісності. Загальна цілісність тексту передбачає аналіз його структурної (зв'язок між реченнями), смислової (єдність теми тексту) та комунікативної (послідовність викладу матеріалу) цілісності. Аналіз кожного з цих аспектів є відповідним кроком текстуального дискурсного аналізу.

Крок перший – виділення структурної цілісності тексту. Досягається за допомогою аналізу слів-маркерів, що представляють собою синоніми одного і того ж поняття X. Представимо набір різних груп, до яких відносяться однотипні поняття послідовністю $X_1, X_2, ..., X_n$. До кожної групи належить k_n однотипних синонімів. Основна тема буде поширюватись по усьому тексту, а тому кількість однотипних синонімів для цієї групи буде значно більша. Пікова величина X буде основною темою тексту, решта груп відзначатиме кожна свою мікротему. Крок другий – аналіз комунікативної цілісності тексту. Досягається шляхом аналізу зв'язності мікротем. Оскільки, при розгортанні основної теми, кожне наступне речення зв'язане з попереднім, тобто додає нову інформацію до попереднього, при цьому частково її повторюючи [16, С. 125-126], позначимо відому інформацію C, а нову D, тоді формула зв'язного послідовного тексту (3) матиме вигляд:

$$\sum_{i=1}^{n}(C_n+D_n)_i, \qquad (3)$$

де n – кількість речень, крок і дорівнює одному реченню.

Кожну групу синонімів X_n слід аналізувати на сліди модифікації засобами стеганографії, порушення умови $D_n \supset C_n \in C_{n-1}$ означає імовірність використання методу заміни синонімів в готовому тексті або штучну генерацію досліджуваного тексту.

Крок третій – аналіз смислової цілісності тексту. Нехай A – множина синонімів основної теми, $B_1, B_2, ..., B_n$ – множини синонімів кожної з мікротем. Множина B_{n-1} не належить до множини B_n. Всі множини B_n належать до множини A. Так $A \supset B_n \in B_{n-1}$ при $\forall B_n \in A$.

Розподіл синонімів основної теми тексту, множини A, поширюється на весь текст, множини $B_1, B_2, ..., B_n$ лише на свою мікротему і на частину наступної, оскільки іде часткове повторення вже відомої попередньої інформації.

Наступним кроком є аналіз множини речень кожної підтеми на наявність слідів стеганографії за допомогою елементів інтенсіональної логіки [17] та семантики Монтегю.

Якщо W – множина можливих світів, від якої залежить I – інтерпретація речення, тоді за умови наявності стегоповідомлення основне смислове навантаження речення втрачається. Отже, розгляду підлягає 2 світи, де w_1 означає, що прихованого підтексту не існує, w_2 означає наявність стегоповідомлення. В даному випадку важливий інтенсіонал пропозиції предикативної групи, оскільки вона володіє значною надлишковістю смислової інформації та може слугувати контейнером для приховування стегоповідомлення. Аналіз цієї структури дасть зрозуміти, наскільки важливу інформацію несе граматична структура, та виявити сліди її модифікації стеганографічними засобами для подальшого видалення структури.

Тоді можна скористатись семантичними правилами, описаними в [17]:

1. Якщо α константа, то $\|\alpha\|^{M,w,g} = I(\alpha)(w)$.
2. Якщо α змінна, то $\|\alpha\|^{M,w,g} = g(\alpha)$.

α – досліджуваний вираз; g – функція з множини змінних всіх типів у відповідну множину значень; M – осмислений вираз. α може бути змінною у випадку, коли належить до світу w_2. В такому випадку $g(\alpha)$ матиме 2 значення, «0» або «1» відповідно до кодування прихованої інформації. Визначити приналежність виразу α до світу w_2 можна шляхом дослідження і порівняння використаних у ньому синонімів з множиною синонімів мікротеми, в якій вони знаходяться, та виявлення порушень умови розподілу мікротем у тексті чи порушенні висхідного розвитку семантики тексту.

Звідси витікає оцінка осмисленості усього тексту шляхом оцінки осмисленості кожного виразу M та простеження їх взаємозалежності у відповідності з правилами побудови зв'язного тексту.

На основі поняття лексичної функції, описаної в [17, С. 263] та основних понять моделі Смисл-Текст, можна допустити, що осмислене твердження матиме вигляд $M = X + Y$, де слову чи сполученню слів X певним чином протиставляється слово чи словосполучення Y, що за певними правилами пов'язано з X за смислом за допомогою одного з типів лексичної функції: лексичної заміни чи лексичних параметрів [17].

Нехай у тексті існує M_n таких тверджень, необхідні умови осмисленості тексту матимуть вигляд:

1. $\forall X_n$ та $\forall Y_n \in B_k$, де B_k – множина синонімів мікротеми k.
2. $B_k \in A$, де A – множина синонімів основної теми.
3. $M_{n_k} \in M_{n_{k+1}}$ та $M_{n_k} \notin M_{n_{k-1}}$, де k – кількість відповідних мікротем.

Загальну осмисленість тексту можна дослідити за формулою (4):

$$Z = \frac{M'_n + T'_n + L'_n + V'_n}{4n} \cdot 100\%, \qquad (4)$$

де M'_n – кількість осмислених тверджень; T'_n – кількість речень, що задовольняють умові 1; L'_n – кількість речень, що задовольняють умові 2; V'_n – кількість речень, що задовольняють умові 3; n – загальна кількість речень. Низький відсоток осмисленості може свідчити про комп'ютерну генерацію тексту.

Оскільки синонімія виникає не лише в лексичному вигляді конкретного слова, слід враховувати це в аналізі та подальшому смисловому стисненні тексту.

Нехай A – кількість активних конструкцій; P – кількість пасивних, тоді A' – кількість активних конструкції, які можна перетворити в пасивні відповідно до наявної у тексті інформації; P' – кількість пасивних конструкції, які можна перетворити в активні; N – кількість речень у тексті. Тоді імовірність приховування інформації шляхом заміни синонімічними пасивними чи активними конструкціями буде визначатися за формулою (5):

$$Z = \frac{(A - A') + (P - P') + \frac{A' + P'}{N}}{A + P} \cdot 100\%, \qquad (5)$$

де на основі отриманих даних можна зробити спробу визначити можливу мету написання тексту. Так, текст буде написаним для потреб стеганографії, якщо виконуватимуться умови, описані в кожному з

кроків аналізу осмисленості тексту з урахуванням формул (1–2).

Наступним кроком є інтертекстуальна перевірка досліджуваної інформації. Таким чином, можна застосовувати метод, описаний в [18]. Відмінність в пошуку не плагіату, а схожого тексту. Таким чином, важливо знайти саме ідентичний не модифікований початковий суцільний текст. Правильно буде скористатися потужними пошуковими системами, як Google. A – множина, що представляє досліджуваний текст, B – множина що представляє знайдений схожий текст. Якщо $|A| = |B|$ та виконується умова $A \subset B \wedge B \subset A$, це означає високу імовірність приховування в ньому стегоповідомлення. В такому випадку необхідно знайти симетричну різницю множин $A \Delta B$.

Етап стиснення тексту. В загальному вигляді етап стиснення текстової інформації на смисловому рівні полягає у позбавленні його смислової надлишковості. Для цього слід звернутися до інтенсіональної логіки, основний підхід якої описано в [17], що можна порівняти з діями над багатовимірними векторами та масивами. Нехай $W_1, W_2, ..., W_n$ – це світи, що утворюються в процесі морфологічно-синтаксичного та дискурсного аналізу і моделювання можливих результатів. $W_1^*, W_2^*, ..., W_n^*$ – множини лексичних одиниць, що належать до світів $W_1, W_2, ..., W_n$ відповідно. Кожен зі світів відповідає за можливе приховування стегоповідомлення відповідним методом стеганографії. Точка дотику всіх світів буде тією незмінною інформацією, яку не можливо модифікувати без втрати змісту тексту та порушення його цілісності та лінійності. Позначимо через T інформацію, що повинна бути збережена після стиснення, тоді вона визначається за формулою (6):

$$T = W_1^* \cap W_2^* \cap ... \cap W_n^*. \quad (6)$$

Виведення загальної формули для проведення скорочення тексту передбачає виконання паралельних етапів видалення, узагальнення і заміни для максимально зниження ризику виникнення помилки.

Спочатку виділяються головні та другорядні члени речення, враховується сталий порядок слів в реченні англійської мови. В загальному випадку речення можна зобразити за формулою (7) згідно із загальними правилами англійської граматики.

$$SENT = (A_{p,t}) + (G) + S + P + (G) + O_{d,i} + A_{w,p,t}, \quad (7)$$

де A – обставина (A_p – обставина місця, A_t – обставина часу, A_w – обставина способу дії); S – підмет; P – присудок; G – означення; O – додаток (O_d – прямий додаток, O_i – непрямий додаток). Кількість кожного зі структурних елементів речення може дорівнювати нулю. У випадку, коли їх більше одного, це говорить про наявність однорідних членів речення або повторів, що повинні бути видалені.

Наступним кроком на етапі видалення є співвідношення смислової інформації кожного елементу речення до загального смислу мікротеми чи основної теми тексту. Якщо будь-який елемент конкретного речення належить до множини синонімів своєї мікротеми $E \in B_n$, проводиться перевірка, чи належить він до вже відомої інформації C. Якщо $E \in C$ при $E \in B_n$ цей елемент слід видаляти.

В загальному випадку, щоб отримати скорочене речення ($SENT'$) на етапі видалення елементів необхідно із загальної моделі речення крім однорідних членів речення та повторів видалити набір елементів ($SENT^*$), в яких існує імовірність використання стеганографічних засобів, і які в той же час не спричинять втрати критичної смислової інформації. Таким чином, скорочення буде відбуватися за принципом (8)

$$SENT' = SENT - SENT^*, \quad (8)$$

де $SENT^* = S^* + P^* + O_{d,i}^*$, а $S^*, P^*, O_{d,i}^*$ – відповідно підмет, присудок та додаток при високій імовірності їх модифікації засобами стеганографії. Так, скорочене речення можна зобразити за формулою (9):

$$SENT' = SENT - A_{w,p,t} - G - SENT^*, \quad (9)$$

або за формулою (10):

$$SENT' = S' + P' + O'_{d,i}, \quad (10)$$

де $S', P', O'_{d,i}$ – підмет, присудок та додаток відповідно при відсутності імовірності їх модифікації. Особливо важливе виконання умови, при якій кожний елемент скороченого речення $E \in Q, E \in R, E \in T$, де Q – множина елементів, в яких існує імовірність використання стеганографічних засобів при морфологічному аналізі; R – при синтаксичному аналізі; T – при дискурсному аналізі. А також при $E \in C$ та $E \in B_n$. Наступний етап стиснення тексту передбачає проведення узагальнення речень.

Якщо $B_1, B_2, ..., B_n$ – множини синонімів n мікротем тексту, тоді для їх узагальнення (на прикладі мікротеми B_1) слід перевіряти умову $b_{1n+1} \in b_{1n}$. Якщо будь-яке речення, яке є елементом множини B_1, належить до групи попереднього речення та за умови відсутності другорядних членів речення в реченні b_{1n+1}, елемент речення $E^{b_{1n}} = E^{b_{1n+1}}$ при $E^{b_{1n}} \in B_1$ виступає ознакою, при якій усю групу речення слід узагальнити.

Схематично речення можна зобразити, як $b_{1n} = E_1^{b_{1n}} + E_2^{b_{1n}} + ... + E_n^{b_{1n}}$, де $E_n^{b_{1n}}$ – елементи

речення b_{1n}. Якщо виконується умова $E_1^{b_{1n}}, E_2^{b_{1n}}, ..., E_n^{b_{1n}} \in Q_{E^{b_{1n}}}^{b_{1n}}$, де $Q_{E^{b_{1n}}}^{b_{1n}}$ – множина семантичних синонімів, що описують один факт у реченні b_{1n}, тоді відбувається узагальнення на основі найбільш підходящого за контекстом елемента $E_n^{b_{1n}}$.

Якщо елемент E досліджуваного речення b_{1n} належить до множини синонімів своєї мікротеми $E \in B_n$, проводиться перевірка, чи належить він до вже відомої інформації C. Якщо $E \in C$ при $E \in B_n$, ці два речення слід узагальнити на основі тотожної інформації.

Для спрощення розрахунків позначимо будь-яке речення b_{1n} чи його елемент $E_n^{b_{1n}}$ як E_n, оскільки для виведення принципової формули узагальнення неважливо, що саме слід узагальнювати, групу елементів конкретного речення чи групу речень. В такому випадку $GEN = E_1 + E_2 + ...E_n$.

Щоб отримати узагальнений уривок тексту (GEN'), необхідно із моделі структурних елементів тексту спершу видалити набір елементів (GEN^*), в яких існує імовірність використання стеганографічних засобів, і які в той же час не спричинять втрати критичної смислової інформації, та набір елементів другого порядку (GEN''), що належать до множини основного смислового елементу (X). Таким чином, при $GEN^* = E_1^* + E_2^* + ... + E_n^*$, де E_n^* – набір з n елементів при високій імовірності їх модифікації засобами стеганографії, а $GEN'' = E_1'' + E_2'' + ... + E_n''$, при $E_n'' \neq X$, де X – перший елемент з одиничних елементів GEN за умови, при якій $X \notin Q, X \notin R, X \notin T$. У випадку, коли X – речення, що належить до множини синонімів своєї мікротеми, а E_n'' – незначні одиничні факти, об'єднані основним смисловим елементом X для його поширення та доповнення, тоді узагальнення буде відбуватися за формулою (11):

$$GEN' = GEN - GEN^* - GEN'', \quad (11)$$

або за формулою (12).

$$GEN' = E_1' + E_2' + ... + E_n'. \quad (12)$$

У випадку, коли основний зміст може бути втрачений при видаленні однорідних членів речення чи неможливо провести узагальнення у зв'язку з відсутністю основного смислового елементу X, якому б підпорядковувались усі інші елементи, необхідно ввести такий елемент штучно та замінити на нього усі інші однорідні елементи. Цей етап стиснення тексту передбачає проведення заміни речень, граматичних структур чи лексичних одиниць.

Якщо $B_1, B_2, ..., B_n$ – множини синонімів n мікротем тексту, тоді для проведення заміні (на прикладі мікротеми B_1) слід перевіряти умову $\forall b_{1n} \in Q^{b_{1n}}$, де $Q^{b_{1n}}$ – множина семантичних синонімів, до яких належить речення. В такому випадку та за умови рівноправності цих синонімів, існує такий елемент b_{1n}', що є синонімом для всієї синонімічної групи. Тоді слід всю групу замінювати на b_{1n}'. Аналогічно і для частин речення (граматичних структур чи однорідних членів). Нехай речення $b_{1n} = E_1^{b_{1n}} + E_2^{b_{1n}} + ... + E_n^{b_{1n}}$, тоді якщо $E_n^{b_{1n}}$ належить до групи лексичних синонімів b_{1n} або є граматичною структурою (прямою мовою), то існує такий елемент E', що об'єднує всю групу $E_n^{b_{1n}}$. В такому випадку група елементів $E_n^{b_{1n}}$ або пряма мова замінюється елементом E'.

Для спрощення розрахунків позначимо будь-яке речення b_{1n} чи його елемент $E_n^{b_{1n}}$ як E_n, оскільки для виведення принципової формули заміни також неважливо, що саме слід замінювати, групу елементів конкретного речення чи групу речень. В такому випадку група структурних елементів тексту, необхідних для заміни $SUB = E_1 + E_2 + ... + E_n$.

Щоб отримати замінений уривок тексту (SUB'), необхідно із структурних елементів тексту спершу видалити набір елементів (SUB^*), в яких існує імовірність використання стеганографічних засобів, і які в той же час не спричинять втрати критичної смислової інформації, та знайти такий елемент Y, який би належав до семантичного значення ряду отриманих елементів, тобто $Y \in (SUB - SUB^*)$, де $SUB^* = E_1^* + E_2^* + ... + E_n^*$ при E_n^* – наборі n елементів з високою імовірністю їх модифікації засобами стеганографії. Таким чином, принцип визначення SUB' можна зобразити з допомогою формули (13):

$$SUB' = \frac{SUB - SUB^*}{SUB - SUB^*} + Y. \quad (13)$$

В результаті формула групи структурних елементів тексту буде виглядати не як ряд, а як один елемент (14):

$$SUB' = Y, \quad (14)$$

за умови, що $Y \notin Q, Y \notin R, Y \notin T$. Оскільки були наведені формули стиснення тексту на кожному етапі, тоді остаточна формула стиснення тексту (15) в цілому матиме наступний вигляд:

$$T' = SENT' \cap GEN' \cap SUB'. \quad (15)$$

Якщо ж розглядати скорочення тексту частинами, що передбачає видалення абзаців чи фрагмен-

тів, то у такому випадку слід аналізувати текст на основі взаємозв'язку його теми з мікротемами та безпосередньо в середні мікротем.

Якщо ряд $B_1, B_2, ..., B_n$ є відповідними мікротемами тексту T та підпорядковується основній темі A, тобто $T = B_1 + B_2 + ... + B_n$, тоді слід аналізувати кожну мікротему на зв'язок з попередньою та основною темою A для надлишковості. В такому випадку, за умови, що $B_n \in A$, $B_n \in B_{n-1}$, слід видаляти мікротему B_n як таку, що продовжує чи конкретизує мікротему B_{n-1}. Позначимо ряд мікротем, що необхідно видалити, як $T' = B'_1 + B'_2 + ... + B'_n$. Звідси слідує формула скорочення тексту (16):

$$T^* = T - T' = B_1^* + B_2^* + ... + B_n^*, \quad (16)$$

за умови, що $\forall B_n^* \notin Q, \forall B_n^* \notin R, \forall B_n^* \notin T$. Наступним кроком скорочення тексту після проведення попередніх етапів є заміна лексичних одиниць у скороченому тексті на еквівалент меншої довжини.

Якщо T' – скорочений текст, що складається з n речень b'_n, кожне з яких є сумою елементів $E_n^{b'_n}$, тобто $b'_n = E'_1 + E'_2 + ... + E'_n$. В такому випадку для будь-якого $E_n^{b'_n}$ може існувати повний лексичний синонім E''_n, кількість символів у якому менше, ніж у $E_n^{b'_n}$, тобто $lenE''_n < lenE_n^{b'_n}$. Тоді слід перевіряти умову відповідності E''_n множині синонімів мікротеми B_n, до якої він належить. Якщо для елемента $E_n^{b'_n}$ існує синонімічний відповідник E''_n, то елемент $E_n^{b'_n}$ при обов'язковому виконанні умови $(E''_n \in B_n) \cap (lenE''_n < lenE_n^{b'_n}) \cap (E''_n \in b'_n)$ замінюється одним з елементів E''_n, що відповідає умові та володіє найменшою довжиною len. Таким чином, стиснений текст з заміною синонімів буде мати вигляд формули (17):

$$b''_n = E''_1 + E''_2 + ... + E''_n. \quad (17)$$

У випадку, коли текст було згенеровано автоматично, його неможливо стиснути звичайними семантичними засобами. Однак, навіть в такому тексті можна визначити A – множину синонімів основної теми та $B_1, B_2, ..., B_n$ – множини синонімів мікротем. Таким чином, якщо не виконуються вищеописані умови осмисленості, стиснення відбувається по принципу видалення усіх мікротем B'_n, що не задовольняють умову $B'_n \in A$. Звідси слідує, що формула стиснення матиме вигляд (18):

$$T^* = T - T' + A = A + B_1^* + B_2^* + ... + B_n^*, \quad (18)$$

де B_n^* – мікротеми, що семантично пов'язані з основною темою A. Отже, якщо мова йде про будь-який текст, що передається, формула стиснення матиме наступний вигляд (19).

$$\begin{cases} (SENT' \cap GEN' \cap SUB') \cap b''_n \\ T - T' + A \end{cases}, \quad (19)$$

оскільки одночасно враховується як осмислений так і неосмислений текст.

Етап модифікації тексту. Оскільки після проведення маніпуляцій з текстом, покликаних його стиснути на смисловому рівні, можуть виникати деякі стилістичні помилки, спричинені видаленням частин речення або частин тексту, тому виникає необхідність модифікувати стиснений текст для виправлення цих помилок, а також для протидії стеганографічним методам довільних інтервалів та частково синтаксичним методам. У загальному вигляді процес модифікації тексту складається з чотирьох послідовних кроків: видалення зайвих інтервалів, виправлення морфологічних помилок, виправлення синтаксичних помилок та до етапів можна віднести вищеописаний заміну синонімів.

Якщо b_n – будь-яке речення у тексті, яке складається з n елементів: $b_n = E_1^{b_n} + E_2^{b_n} + ... + E_n^{b_n}$, тоді за умови, що кожний з елементів $E_n^{b_n}$ є існуючим словом, це дозволяє зробити висновок, що між $E_n^{b_n}$ та $E_{n-1}^{b_n}$ не пропущений інтервал, тоді усі $E_n^{b_n}$ повинні бути поєднані між собою інтервалом F. Завжди повинна виконуватись умова, що об'єднуючий інтервал $F_n = 1$. Наступним кроком є дослідження тексту на наявність додаткових інтервалів між реченнями по аналогії з елементами речення.

Наступним кроком модифікації тексту є виправлення морфологічних помилок. Для цього необхідно звернутися до етапу морфологічно-синтаксичного аналізу, побудувати модель досліджуваного речення та порівняти її з початковою моделлю і провести перевірку на відповідність правилам граматики та правопису шляхом перевірки за словником. Якщо E_n – елемент досліджуваного речення, при наявності ознак часу, тривалості, порівняння, множини в елементі E_n, або ж у випадку, якщо E_n є допоміжним дієсловом, тобто так чи інакше вказує на використання відповідної форми елемента E, який залежить від елементу E_n та співвідноситься з ним за певними граматичними правилами t, проводиться перевірка виконання умови $E + E_n \Leftrightarrow E^t + E_n$. У випадку, коли умова не виконується, E приводиться до вигляду E^t.

Далі проводиться виправлення синтаксичних помилок по аналогії з кроком морфологічної модифікації тексту. Визначається час дієслова, знаходяться ознаки тривалості чи множини. Якщо в стру-

ктурному елементі речення E_n, або ж у випадку, якщо E_n є допоміжним дієсловом to be у тій чи іншій своїй формі, присутні ознаки, що вказують на необхідність використання відповідної форми елемента E, який залежить від елементу E_n та співвідноситься з ним за певними синтаксичними правилами s, проводиться перевірка умови $E + E_n \Leftrightarrow E^s + E_n$. У випадку, коли умова не виконується, E прирівнюється до E^s.

Наступним кроком модифікації тексту можна вважати заміну слів синонімами меншої довжини. Якщо елемент речення E_n належить до множини елементів E_n^*, в яких існує висока імовірність використання стеганографічних засобів, що була виявлена на етапі морфологічно-синтаксичного та дискурсного аналізу, тоді виконання частини умови $(E_n'' \in B_n) \cap (lenE_n'' < lenE_n^{b_n'}) \cap (E_n'' \in b_n')$, а саме $lenE_n'' < lenE_n^{b_n'}$ не є обов'язковим. Перевіряється приналежність E_n'' до групи синонімів B_n. Таким чином при виконанні умови $(E_n'' \in B_n) \cap (E_n'' \in b_n')$ E_n замінюється на E_n''. Перевага все ж віддається синоніму з $lenE_n'' < lenE_n^{b_n'}$, проте, якщо такого не існує, то із ряду обирається синонім з найменшою кількістю символів при обов'язковому виконанні умови $(E_n'' \in B_n) \cap (E_n'' \in b_n')$ навіть якщо $lenE_n'' \geq lenE_n^{b_n'}$.

Після внесення всіх змін проводяться відповідні послідовні дії по формуванню остаточного тексту.

Формування тексту після внесення змін. Останнім етапом методу є формування результуючого тексту з врахуванням особливостей стилістики синтаксичної структури оригінального тексту та семантичної структури його дискурсу. В загальному вигляді алгоритм складається з трьох послідовних кроків: формування морфологічно-синтаксичної структури, реалізація семантичної відповідності, опис усіх внесених модифікацій та видалень.

Якщо речення в початковому тексті T до проведення маніпуляцій по скороченню та модифікації було двоскладним $b_n^T = S + P + E_n$, де S – підмет, P – присудок, E_n – другорядні члени речення, а $b_n^{T'}$ стало після стиснення односкладним, тобто $b_n^{T'} = S + E_n$ або $b_n^{T'} = P + E_n$, тоді його слід об'єднати з $b_{n-1}^{T'}$, якщо виконується умова $(b_n^T \cap b_{n-1}^{T'}) \in B$, де B – відповідна мікротема. Позначимо елементи речення $b_{n-1}^{T'} = S' + P' + E_n'$, тоді нове сформоване речення $b_{n-1}' = b_{n-1}^{T'} - b_n^{T'}$ і розраховується з формулою (20):

$$\begin{cases} b_{n-1}' = S + P' + E_n + E_n'; \\ b_{n-1}' = S' + P + E_n + E_n'; \\ b_{n-1}' = S + P + E_n + E_n'. \end{cases} \quad (20)$$

За умови відсутності C_n, відомого у реченні n, речення n та n–1 об'єднуються і приводяться до вигляду $C_{n-1} + D_{n-1} + D_n$. Це означає, що виконується етап остаточного формування морфологічно-синтаксичної структури вихідного обробленого тексту. Після приведення синтаксичної структури тексту до осмисленого вигляду слідує реалізація семантичної відповідності. Маніпуляції з реченням забезпечують комунікативну цілісність тексту. В свою чергу структурна і смислова цілісність досягається шляхом формування тексту в цілому, відтворення мікротем у відповідній послідовності.

Нехай $B_1, B_2, ..., B_n$ – послідовність множин мікротем початкового тексту; $B_1', B_2', ..., B_n'$ – послідовність множин мікротем стисненого і модифікованого тексту. Усі мікротеми повинні бути виставлені згідно з послідовністю початкового тексту за умови, що $B_{n-1} \times B_n = B_{n+1}$, тобто кожна множина синонімів мікротеми послідовно семантично наслідує попередні множини мікротем, завдяки чому наслідується їх семантика.

Проте, коли множини B_n та B_{n+1} рівноправні та паралельно розкривають смисл множини B_{n-1}, в такому разі послідовність перестановок B_n та B_{n+1} неважлива. В результаті, забезпечується необхідні умови структурної цілісності результуючого тексту.

Також, при послідовному розподілу, частина семантичної інформації з B_n повинна поширюватись на B_{n+1}, за умови, що $B_n \in A$ та $B_{n+1} \in A$, де A – множина синонімів основної теми тексту. Таким чином забезпечується смислова цілісність результуючого тексту. В результаті штучного перемішування мікротем, порушується структура можливого кодування секретного повідомлення. Наступним є крок опису модифікацій. Таким чином, якщо E_n^T – один з n елементів початкового тексту T, то при $E_n^T \in Q, E_n^T \in R, E_n^T \in T$, де Q – множина елементів, в яких існує імовірність використання стеганографічних засобів при морфологічному аналізі, R – при синтаксичному аналізі, T – при дискурсному аналізі застосовується форматування першого типу. Форматування другого типу застосовується до елементу E_n^T за умови, що $E_n^T \notin T'$, якщо $E_n^T \in J, E_n^T \in H, E_n^T \in F$, де J – множина елементів, що були видалені, H – узагальнені, F – замінені на

етапі стиснення тексту. Форматування третього типу застосовується за умови, що $E_n^T \notin T'$, при $E_n^T \in L, E_n^T \in C, E_n^T \in S$, де L – множина елементів, які були модифіковані на етапі виправлення морфологічних помилок, C – синтаксичних помилок, S – при заміні синонімами.

В загальному вигляді принцип побудови результуючого тексту можна зобразити за допомогою формули (21):

$$T' = \sum_{i=1}^{n} (\sum_{i=1}^{n} b_i')^{B_i}, \quad (21)$$

на основі чого формується остаточний текст після внесення усіх змін і маніпуляцій.

Висновки

Було розроблено метод семантичного стиснення текстової інформації для протидії комп'ютерній лінгвістичній стеганографії, що складається з п'яти етапів. Етап автоматизованого лінгвістичного аналізу тексту забезпечує визначення імовірності модифікації усього тексту, як і окремих речень з різною відсотковою імовірністю завдяки синтаксичному дослідженню. Етап оцінки осмисленості тексту та виділення основної думки перш за все передбачає визначення загальної цілісності тексту з подальшим аналізом множини речень кожної підтеми на наявність слідів стеганографії за допомогою елементів інтенсіональної логіки та семантики Монтегю, оцінкою осмисленості виразів і усього тексту, виділенням мети його написання, порівнянням з аналогами в мережі. Етап стиснення тексту зводиться до знаходження варіанту цього тексту, що лежить на перетині паралельних між собою етапів видалення, узагальнення та заміни та видалення надлишковості за умови, що текст володіє ознаками структурованості, зв'язності, відповідає необхідним умовам осмисленого тексту та видалення неструктурованих даних, що не пов'язані з головною темою тексту у випадку, коли текст є штучно згенерованим. На етапі модифікації тексту виправляються стилістичні помилки, спричинені видаленням частин речення або частин тексту. Крім того, цей етап зможе забезпечити видалення залишкових слідів використання засобів стеганографії у вигляді частин стегоповідомлення, прихованих у важливих елементах тексту, які неможливо видалити у зв'язку з їх ключовою позицією у семантичній структурі та забезпечує додаткове семантичне стиснення тексту на 2–3%. Етап формування тексту після внесення змін забезпечує інформаційно-структурні якості тексту та встановлює відповідність між семантичною структурою сформованого та початкового тексту. Таким чином, можна стверджувати, що метод є основою для автоматизації атаки на лінгвістичну стегосистему шляхом семантичного стиснення та подальшої програмної реалізації. Функціональні можливості, що надає метод, а саме стиснення як окремих речень так и текстів великого об'єму, широкий спектр стегоаналізу, виявлення штучно згенерованих текстів, врахування семантики та можливість видалення стегоповідомлення при стисненні відкривають широкий спектр для подальшого практичного застосування методу у комплексних комп'ютерних, стегоаналітичних системах та для атак на лінгвістичну стегосистему, основану на використанні природньої (англійської) мови.

Список літератури

1. Жмакин М.О. Стеганография и перспективы ее применения в печатних документах / М.О. Жмакин // Безопасность информационных технологий. – 2010. – №3. – С. 74-77.

2. Chen Z. Effective Linguistic Steganography Detection / Z. Chen, L. Huang, Z. Yu, X. Zhao, X. Zhao // 8th International Conference on Computer and Information Technology Workshops. – 08-11 July, 2008. – Sidney, Australia. – P. 224-229.

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

4. Meng P. STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography / P. Meng, L. Hang, Z. Chen, Y. Hu, W. Yang // 12th International Conference «Information Hiding», June 28-30, 2010. – Vol. 6387. – Calgary, Canada. – P. 208-220.

5. Chen Z. Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words / Z. Chen, L. Huang, Z. Yu, W. Yang, L. Li, X. Zheng, X. Zhao // 10th International Workshop «Information Hiding», May 19-21, 2008. – Vol. 5284. – Santa Barbara, USA. – P. 224-235.

6. Taskiran C.M. Attacks on lexical natural language steganography systems / C.M. Taskiran, U. Topkara, M. Topkara, E. J. Delp // Proceedings of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VIII, 15 February 2006. – Vol. 6072. – San Jose, USA. – P. 97-105.

7. Утилиты лингвистического анализа текста (морфология, синтаксис) [Електронний ресурс] // Программы лингвистического анализа и обработки текста. – Режим доступу до ресурсу: http://www.asknet.ru/analytics/programms.htm.

8. Большакова И.Е. Автоматический анализ дискурсивной структуры научного текста / И.Е. Большакова, Н.В. Баева // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2004. – Тверь, Россия, 2-7 июня, 2004. – С. 68-73.

9. Erkan G. LexRank: graph-based lexical centrality as salience in text summarization / G. Erkan, D.R. Radev // Journal of Artificial Intelligence Research. – 2004. – Vol. 22, Issue 1. – P. 457-479.

10. Paulus R. A Deep Reinforced Model for Abstractive Summarization [Електронний ресурс] / R. Paulus, C. Xiong, R. Socher // arXiv preprint arXiv:1705.04304. – 2017. – Режим доступу до ресурсу: https://arxiv.org/abs/1705.04304.

11. Radev D.R. Centroid-based summarization of multiple documents / D.R. Radev, H. Jing, M. Stys, D. Tam // Information Processing & Management. –2004. – № 40(6). – P. 919-938.

12. Бабина О.И. Лингвистическая стеганография: современные подходы. Часть 2 / О.И. Бабина // Вестник ЮУрГУ. Серия: Лингвистика. – 2015. – №4. – С. 49-55. https://doi.org/10.14529/ling150410.

13. Евсеев С.П. Использование ущербных кодов в крипто-кодовых системах / С.П. Евсеев // Системи обробки інформації. – 2017. – № 5 (151). – С. 109-121. https://doi.org/10.30748/soi.2017.151.15.

14. Федотова-Півень І.М. Шляхи задоволення потреб сучасної кібербезпеки в рамках протидії методам комп'ютерної лінгвістичної стеганографії / І.М. Федотова-Півень, Я.В. Тарасенко // Безпека інформації. – 2017. – №23(3). – С. 190-196. https://doi.org/10.18372/2225-5036.23.12093.

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

16. Глухов В.П. Психолингвистика: учебник и практикум для академического бакалавриата / В.П. Глухов. – М.: Юрайт, 2017. – 361 с.

17. Осипов Г.С. Методы искусственного интеллекта / Г.С. Осипов. – М.: ФИЗМАТЛИТ, 2011. – 296 с.

18. Крутояров Д. В. Автоматизированная система поиска заимствований в электронных изданиях, опубликованных в сети Интернет: дис. … к. т. н.: 05.13.06 / Д.В. Крутояров. – М., 2006. – 191 с.

References

1. Zhmakin, M.O. (2010), "Steganografiya i perspektivyi ee primeneniya v pechatnih dokumentah" [Steganography and prospects for its use in printed documents], S*ecurity of Information Technologies*, No. 3, pp. 74-77.

2. Chen, Z., Huang, L., Yu, Z., Zhao, X. and Zhao, X. (2008), Effective Linguistic Steganography Detection, *8th International Conference on Computer and Information Technology Workshops*, July 08-11, Sidney, Australia, pp.224-229.

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

4. Meng, P., Hang, L., Chen, Z., Hu, Y. and Yang, W. (2010), STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography, *12th International Conference «Information Hiding»*, June 28-30, Calgary, Canada, pp. 208-220.

5. Chen, Z., Huang, L., Yu, Z., Yang, W., Li, L., Zheng, X. and Zhao, X. (2008), Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words, *10th International Workshop «Information Hiding»*, May 19-21, Santa Barbara, USA, pp. 224-235.

6. Taskiran, C.M., Topkara, U., Topkara, M. and Delp, E.J. (2006), Attacks on lexical natural language steganography systems, *Proceedings of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VIII*, 15 February, San Jose, USA, pp. 97-105.

7. Utilities of the textual linguistic analysis (morphology, syntax) "*Programmy lingvisticheskogo analiza i obrabotki teksta*" [*The programs of linguistic analysis and text processing*], www.asknet.ru/analytics/programms.htm (accessed 13 April 2018).

8. Bolshakova, I.E. (2004), "Avtomaticheskiy analiz diskursivnoy strukturyi nauchnogo teksta" [The discursive structure of scientific text automatic analysis], *Computer linguistics and intellectual technologies: Proceedings of the international conference Dialogue'2004*, 2-7 June, Tver, Russia, pp. 68-73.

9. Erkan, G. and Radev, D.R. (2004), LexRank: graph-based lexical centrality as salience in text summarization, *Journal of Artificial Intelligence Research*, Vol. 22, Issue 1, pp. 457-479.

10. Paulus, R., Xiong, C. and Socher, R. (2017), A Deep Reinforced Model for Abstractive Summarization, *arXiv preprint arXiv:1705.04304*, https://arxiv.org/abs/1705.04304 (accessed 19 April 2018).

11. Radev, D.R., Jing, H., Stys, M. and Tam, D. (2004), Centroid-based summarization of multiple documents, *Information Processing & Management*, No. 40(6), pp. 919-938.

12. Babina, O.I. (2015), "Lingvisticheskaya steganografiya: sovremennyie podhodyi. Chast 2" [Linguistic steganography: modern approaches. Part 2], *the South Ural State University Newsletter. Series: Linguistics*, No. 4, pp. 49-55. https://doi.org/10.14529/ling150410.

13. Yevseiev, S. (2017), "Ispolzovanie uscherbnyih kodov v kripto-kodovyih sistemah" [The use of damaged codes in crypto code systems], *Information Processing Systems*, No. 5 (151), pp. 109-121. https://doi.org/10.30748/soi.2017.151.15.

14. Fedotova-Piven, I.M. and Tarasenko, Ya.V. (2017), "Shliakhy zadovolennia potreb suchasnoi kiberbezpeky v ramkakh protydii metodam kompiuternoi linhvistychnoi stehanohrafii" [The ways of meeting the needs of modern cybersecurity in context of countering the methods of computer linguistic steganography], *Ukrainian Scientific Journal of Information Security*, No. 23(3), pp. 190-196. https://doi.org/10.18372/2225-5036.23.12093.

15. Andreev, A.M. (1999), "Veroyatnostnyiy sintaksicheskiy analizator dlya informatsionno-poiskovoy sistemyi" [A probabilistic parser for information retrieval system], *Computer Chronicle*, No. 1, pp. 37-85.

16. Gluhov, V.P. (2017), "*Psiholingvistika: uchebnik i praktikum dlya akademicheskogo bakalavriata*" [*Psycholinguistics: a tutorial and a workshop for the academic undergraduate*], Urait, Moscow, 361 p.

17. Osipov, G.S. (2011), "*Metodyi iskusstvennogo intellekta*" [*The methods of artificial intelligence*], Fizmatlit, Moscow, 296 p.

18. Krutoyarov, D.V. (2006), "*Avtomatizirovannaya sistema poiska zaimstvovaniy v elektronnyih izdaniyah, opublikovannyih v seti Internet: dissertation*" [*The automated system for searching the borrowings in the Internet electronic publications: dissertation*], Moscow, 191 p.

Надійшла до редколегії 19.07.2018
Схвалена до друку 21.08.2018

Відомості про авторів:

Тарасенко Ярослав Володимирович
аспірант кафедри Черкаського державного технологічного університету,
Черкаси, Україна
https://orcid.org/0000-0002-5902-8628

Півень Олег Борисович
кандидат фізико-математичних наук доцент професор кафедри Черкаського державного технологічного університету,
Черкаси, Україна
https://orcid.org/0000-0002-3984-6439

Федотова-Півень Ірина Миколаївна
кандидат технічних наук доцент завідувач кафедри Черкаського державного технологічного університету,
Черкаси, Україна
https://orcid.org/0000-0002-0512-6118

Information about the authors:

Yaroslav Tarasenko
Postgraduate Student of Department of Cherkasy State Technological University,
Cherkasy, Ukraine
https://orcid.org/0000-0002-5902-8628

Oleg Piven
Candidate of Physics and Mathematics Associate Professor Professor of Department of Cherkasy State Technological University,
Cherkasy, Ukraine
https://orcid.org/0000-0002-3984-6439

Iryna Fedotova-Piven
Candidate of Technical Sciences Associate Professor Head of Department of Cherkasy State Technological University,
Cherkasy, Ukraine
https://orcid.org/0000-0002-0512-6118

МЕТОД СЕМАНТИЧЕСКОГО СЖАТИЯ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ПРОТИВОДЕЙСТВИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИЧЕСКОЙ СТЕГАНОГРАФИИ

Я.В. Тарасенко, О.Б. Пивень, И.Н. Федотова-Пивень

В статье приводится разработка метода семантического сжатия текстовой информации для противодействия компьютерной лингвистической стеганографии, состоящий из пяти последовательных этапов (автоматизированного лингвистического анализа, оценки осмысленности текста и выделение основной мысли, сжатия, модификации и формирования окончательного текста после внесения изменений) на основе дискурсного анализа, который отличается адаптацией методов реферирования текста для учета возможного использования средств стеганографии и проведения атаки на лингвистическую стегосистему путем семантического сжатия как осмысленных, так и искусственно сгенерированных текстов на основе комплексного подхода и учета широкого спектра средств стеганографии.

Ключевые слова: *компьютерная лингвистическая стеганография, семантическое сжатие текста, текстовый стегоанализ, автоматизированный стегоанализ, метод сжатия текста, определения осмысленности текста, атака на лингвистическую стегосистему.*

METHOD OF THE TEXTUAL INFORMATION SEMANTIC COMPRESSION FOR COUNTERACTING COMPUTER LINGUISTIC STEGANOGRAPHY

Ya. Tarasenko, O. Piven, I. Fedotova-Piven

There is a need to develop methods that can provide an effective attack on the linguistic stegosystem in modern realities. Attack based on semantic compression is the most effective mean of counteracting any method of textual steganography for the purpose of removing the stegomessages. However, there is a problem associated with the computerization of the process. The article describes the method of the textual information semantic compression for counteracting computer linguistic steganography, consisting of five consecutive stages (automated linguistic analysis of the text, evaluating its comprehension and allocation of the basic meaning, compression, modification and formation of the final text after making changes) based on the discursive analysis, which is determined by the adaptation of the summarization methods to take into account the possible steganography means to the attack the linguistic stegosystem by semantic compressing of both, meaningful and automatically generated texts. The method is based on usage of an integrated approach and takes into account a wide range of steganography methods. It was developed using probability theory, mathematical statistics and simulation, set theory, the method of solving tasks using Euler circles, morphological (the logical multiplication method), syntactic (probabilistic-statistical method) and discourse analysis (textual-intertextual approach); summarization methods and comprehensions writing ways, elements of intensional logic and formal grammar. The proposed method is the basis for automating the attack on the linguistic stegosystem by semantic compression and further program implementation. The functionality provided by the method, namely the compression of sentences and large texts, a wide range of research, the detection of automatically generated texts, taking into account the semantics and the possibility of removing the stegomessage during the compression, open a wide range of further practical application of the method in complex computer steganographic and steganoanalytic systems and attacks on a linguistic stegosystem based on the use of natural (English) language in its basis.

Keywords: *computer linguistic steganography, semantic compression of the text, textual steganalysis, automated steganalysis, method of textual compression, determination of the textual meaningfulness, attack on the linguistic stegosystem.*