СЕМАНТИЧЕСКИЙ И ЛЕКСИКО-СИНТАКСИЧЕСКИЙ ПОДХОДЫ В СТЕГАНОГРАФИИ (НА МАТЕРИАЛЕ ПИСЬМЕННЫХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ)

А.В. Джунковский
Московский Государственный Лингвистический Университет
Москва

Развитие компьютерных технологий сделало криптографические методы защиты информации относительно ненадежными. В этих условиях стеганографические методы вновь обретают высокую актуальность.

Как правило, все методы лингвистической стеганографии основаны на сокрытии информации в речи, будь то устная или письменная разновидность. При этом не имеет значения, воспроизводится ли сообщение (аудиозапись, текст) или нет (единичная реплика). Наиболее значимым является неочевидность истинного сообщения, которое должно быть понятно лишь конечному адресату. Все остальное представляет собой осмысленный и связный поток информационного шума, который, в силу особенностей стеганографии, может нести полезную информацию, не являющуюся релевантной для истинных целей отправителя и получателя сообщения.

Другими словами, осмысленное сообщение является лишь контейнером, в рамках которого коммуникативная функция достигается не эксплицитным образом с помощью прямой интерпретации смысла сообщения, а с помощью использования методики, известной лишь отправителю и получателю.

В случае, если сообщение передается устно, оно оказывается более защищенным от автоматического стегооанализа в связи с необходимостью создания абсолютно дикторонезависимой системы, которая могла бы с крайне низким уровнем погрешности воспринимать звучащую речь. Те же проблемы возникают в тех случаях, когда энигмограмма обладает рукописным видом. Даже в случае ручной оцифровки рукописного текста, стегоаналитик столкнется с утерей огромного пласта информации, которая могла нести скрытое значение (наклон почерка, переходы между буквами, давление), что делает автоматизированный анализ таких энигмограмм весьма ненадежным методом криптоанализа [1].

Стегоанализ как научная дисциплина, несмотря на давнюю историю стеганографии, находится на ранних этапах своего развития. Существует лишь малое число работ, посвященных этой тематике. Почти все из них посвящены нелингвистическим методам стегоанализа.

Наиболее развиты методы стегоанализа цифровой информации. Так, разработана определенная методика стегоаналитической атаки на контейнер, представляющий собой изображения, звуковые файлы, видеозаписи и метаданные. Все названные методы ограничены по своему применению и выполняются, в большинстве случаев, с применением компьютерной техники. Главным инструментом стегоанализа в этих случаях является статистический анализ параметров аудиофайла, изображения или метаданных. Иными словами, стегоаналитики анализируют формальные характеристики названных типов данных на предмет наличия статистических погрешностей [3]. Эти аномалии позднее интерпретируются вручную [4].

Иными словами, если такая интерпретация, т.е. стегоаналитическая атака оказывается безуспешной, у стегоаналитиков в наличии оказываются лишь статистические данные, на основе которых можно сделать вывод, что файл с определенной вероятностью подвергался изменениям, которые, в свою очередь, с некоторой вероятностью могли носить характер стеганографического кодирования информации. В случаях, когда обнаруженные данные оказываются повторно зашифрованы криптографическими методами, шанс вскрытия шифра оказывается крайне низким [2].

В случаях, когда речь идет о лингвистических методах стеганографии, расшифровка сообщения становится еще сложнее.

Рассмотрим семантический и лексико-синтаксический методы стеганографического сокрытия информации для письменных текстов русского языка. Семантические методы подразумевают использование системы «триггеров», минимальных искажений текста, которые в себе содержат крупное закодированное сообщение, содержание которого известно сторонам заранее. В этих случаях стеганограмма служит лишь для того, чтобы дать сигнал о том, что заложенное заранее сообщение должно быть воспринято получателем энигмограммы [5].

В случае с лексико-семантическим уровнем, речь идет о кодировании единичных символов, слов или предложений и передаче их в около-кодовой форме. В этом случае непосредственное сообщение полностью представлено в энигмограмме, однако сокрыто с помощью установленного кода. В этих случаях стеганограмма по своей природе оказывается близка к криптограмме.

Проведенный нами анализ показал, что существует три принципиально различных типа контейнеров для целей стеганографии: метаэлементы текста, лингвистическое содержание, а также контекст. Приведем созданную нами типологию возможных контейнеров, разделенных по данному критерию.

1. Внедрение стего в элементы оформления и форматирования текста (искажение метаэлементов текста)
   a. использование пробелов между словами;
   b. использование пробелов между буквами;
   c. использование расположения и размера красных строк;
   d. использование расположения верхних и нижних границ отдельных слов и букв;
   e. использование размера шрифта;
   f. использование цвета шрифта;
   g. использование типа шрифта;
   h. использование размеров межстрочных интервалов;
   i. использование наклона всего текста
   j. использование наклона отдельных элементов текста;
   k. модификация написания отдельных графем;
   l. модификация переходных элементов между графемами (применительно к рукописному тексту).
2. Внедрение стего в лингвистическое содержание текстового сообщения (искажение лингвистического содержания)
   a. внедрение эрративов правописания;
   b. искажение на морфемном уровне;
   c. искажение на лексическом уровне;
   d. искажение на синтаксическом уровне;
   e. стилистическое искажение;
   f. использование анахронизмов и неологизмов.
3. Внедрение стего на базе экстралингвистических особенностей текста (использование контекста для целей создания стего)
   a. смена стиля или упомянутых элементов содержания по сравнению с предыдущими текстами, направленными этому адресанту;
   b. искаженное упоминание реалий текущего времени;
   c. несоблюдение корректного использования норм вежливости.

Первый тип контейнеров предполагает использование аномалий в оформлении и форматирования текста. Второй — использование самого лингвистического содержания. Наконец, третий — контекст и экстралингвистические особенности текста.

Семантические методы по своему применению оптимальны для передачи крайне важных единичных сообщений или небольших групп сообщений, т.к. их обнаружение крайне сложно, а расшифровка практически невозможна. В свою очередь, лексико-синтаксические методы стеганографии подходят для продолжительного обмена скрытыми сообщениями, однако значительно более уязвимы, т.к. раскрытие алгоритма автоматически нарушает конфиденциальность всех сообщений между адресатом и адресантом.

ЛИТЕРАТУРА

1. Bennet K. Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text // CERIAS Tech Report. 2013. № 4. Pp. 1-30.
2. Chen Z., Huang L., Meng P. Blind Linguistic Steganalysis against Translation Based Steganography // Lecture Notes in Computer Science. 2010. Vol. 6526. Pp. 251-265.
3. Wayner P. Disappearing Cryptography: Information Hiding & Steganography and Watermarking. San Francisco: Morgan Kaufmann, 2009. Pp. 340-345.
4. Xiang L. Sun X., Luo G. Linguistic Steganalysis using the features derived from Synonym Frequency // Multimedia Tools and Applications. 2014. № 3. Pp. 1893-1911.
5. Г.В. Лукьянов. Основы кодирования и криптографического преобразования информации. Учебное пособие. Москва: ООО «Сарма», 2005. 128 с.