Ю.П.Гульчак

ЛІНГВІСТИЧНІ МЕТОДИ УЩІЛЬНЕННЯ І ВІДНОВЛЕННЯ ІНФОРМАЦІЇ

Лінгвістичні та психолінгвістичні особливості скорочення текстів (ущільнення) дозволяють ефективніше використовувати і шифрування даних. Актуальність проблеми підсилюють і перевантаження каналів зв'язку, і інтенсивність потоків даних.

Автором запропонований варіант ущільнення даних, який використовує природну збитковість текстових повідомлень. Відомо, що збитковість європейських мов наближено однакова і складає біля 60%. Це значить, що при втраті половини тексту є можливість відновити його смисл.

В роботі розглянуті і проаналізовані кілька підходів до ущільнення, кодування і відновлення даних. Перший з них передбачає нові принципи побудови алфавітів. Для передачі текстових матеріалів використовуються алфавіти з різною кількістю символів (латинський, кирилиця тощо); розширені алфавіти (типу ієрогліфів з тисячами знаків, які відповідають окремим словам і словосполученням); скорочені алфавіти – коли одні і ті ж символи мають різні значення в залежності від місця в слові, поєднання з іншими знаками мови тощо. Враховуючи це

запропоновано штучно скоротити а кількість знаків українського алфавіту, видаливши голосні літери.

Другий спосіб враховує, що існує доцільне з точки зору розуміння тексту скорочення окремих типових слів і виразів. Прикладом таких текстів є студентські конспекти. Такий метод актуальний у зв'язку з підвищеною зацікавленістю проблемами дистанційного навчання.

Для реалізації першого способу була висунута ідея створення скороченого алфавіту з обмеженим набором літер, але потужним словником, що дозволили б суттєво зменшити об'єм переданої інформації і одночасно застосувати відомі алгоритми ущільнення даних. Даний метод дає можливість без шкоди для розуміння повідомлення видаляти частину інформації, не передавати її, але повністю відтворити при прийомі адресатом.

Як приклад вирішення такої задачі були досліджені повідомлення, які формувались лише з приголосних літер українського алфавіту. Аналіз багатьох прикладних текстів включаючи і технічні показав, що ступінь їх розуміння фахівцем навіть без спеціальних словників сягає 95% і більше. Видалення голосних літер при збереженні розділових знаків і пробілів скорочує об'єм повідомлення в середньому на 25%.

Якщо ентропію (кількість інформації на символ повідомлення) українського тексту прийняти за 100%, то ентропія повідомлення з приголосних і пробілу становить біля 71%, а приголосних, пробілу і знаків пунктуації -78% і зменшується при врахуванні кореляції сусідніх літер.

Для більш точного відтворення скороченого повідомлення був створений словник на основі аналізу багатьох повідомлень і використання певних статистичних закономірностей появи символів чи слів у повідомленні. Такий словник поповнюється з кожним новим повідомленням. У випадку, якщо інформація стосується лише певної конкретної галузі, можна створити спеціалізований словник і отримати можливість ще більше скоротити об'єм інформації, що передається.