

Efficient Lexical Encoding of Natural Language

Heers, Brian (School: InTech Collegiate High School)

This project detailed the development of NLC (Natural Language Compression), a text encoder optimized for natural language (NL). Efficient NL encoding is most valuable for satellite-based text communication, where the cost of transmission exceeds the cost of computing (encoding/decoding resources), and research applications where reduced bandwidth is necessary to ensure data reception. In contrast to current methods, which encode on a character-basis, NLC operated on a lexeme-basis. A lexeme was defined as a word, punctuation mark, or multiple words often used together ("such as"). NLC assigned identifiers to lexemes, which were variably encoded to have shorter bit-lengths based on how frequently the lexeme is used in NL (according to a static dictionary predefined on both ends). NLC combines this static dictionary with standard usages of punctuation and spaces to implicitly encode information from context. On average, when used to encode NL, NLC returned binary 3.37 times shorter than ASCII, a current standard for encoding NL, and 2.77 times shorter than LZW, a current standard for text compression. On tests individually performed on over 200,000 sentences from the web (originating from the same source as the dictionary, adequately representing NLC for general purposes), 97.4% were encoded over twice as efficient as ASCII and 72% of them, over three times as efficient. These findings demonstrated that NLC could significantly reduce transmission costs in bandwidth-limited environments, especially for short (sentence-length) texts. Future work will explore dictionary flexibility, especially with multilingual applications.

Awards Won: