Better text compression from fewer lexical n-grams

Tony C. Smith and Michelle Lorenz
Computer Science, University of Waikato, Hamilton, New Zealand.
tcs@cs.waikato.ac.nz

Word-based context models for text compression have the capacity to outperform more simple character-based models, but are generally unattractive because of inherent problems with exponential model growth and corresponding data sparseness. These ill-effects can be mitigated in an adaptive lossless compression scheme by modeling syntactic and semantic lexical dependencies independently.

If one divides the vocabulary of a language into two broad non-overlapping classes—a set T comprised solely of thematic/semantic terms (e.g. nouns, verbs, adjectives, etc.) and a set F comprised of functional/grammatical terms (e.g. determiners, auxiliaries, prepositions, etc.)—then language can be viewed as the interlacing of two lexical streams: a semantic sequence and a grammatical sequence. Two words are said to be *super-adjacent* if they are next to each other in one of the two streams. A lexical bigram model of super-adjacent terms enhances the already high mutual information for close proximity semantic words and preserves the strong syntactic dependencies exhibited in patterns of grammatical words, leading to better compression under a PPM-style arithmetic coding scheme.

A conventional bigram model with a vocabulary $V = T + F$ has $(T + F)^2$ bigrams in a comprehensive model. In comparison, a super-adjacency model with separate semantic and functional contexts has only $T^2 + F^2$ bigrams, helping reduce the problem of data sparseness. These gains are admittedly quite small, but distinguishing between T and F creates the opportunity for considerable improvement. Given that semantic dependencies are generally assumed between thematic base forms rather than explicit words—such that the three separate lexical relationships manifest as "girl eats", "girls eating" and "girls eat" are effectively captured in the one semantic relationship "girl eat"—T can be significantly reduced through desuffixion of regularly inflected words. Important syntactic information (such as agreement) embodied in inflectional suffixes is preserved by treating those suffixes as free-standing functional morphemes in the grammatical stream—a treatment that is consistent with modern linguistic theory. Desuffixion thus effectively reduces T by nearly a half, and adds but four terms to F (i.e. the suffixes *-s*, *-ing*, *-ed* and *-'s*), giving a total model size of about $(T/2)^2 + (F + 4)^2 \ll V^2$. The result is sufficiently rapid convergence, and ultimately better overall compression from a substantially smaller model.

Results from experiments with the Brown Corpus show that the super-adjacency bigram model delivers almost 8.5% better compression than conventional word-based bigrams, yet does so from fewer than two-thirds the number of observed contexts. Moreover, compression and model compactness continue to improve over equivalent conventional models as context length increases. Similar results are observed across a wide range of test corpora, indicating the approach is robust for English text in general. The only language specific component of the system is its heuristic desuffixion routine, and it is conjectured (and presently being tested) that comparable performance can be achieved for any non-agglutinating language given procedures for appropriate morphosyntactic analysis. It is further conjectured that the general approach could work well for any data that exhibits super-adjacent dependencies.

Proceedings of the Data Compression Conference (DCC'01)
1068-0314/01 $10.00 © 2001 **IEEE**