IAA-99-IAA.9.1.08

Language in Signals: the Detection of Generic Species-Independent Intelligent Language Features in Symbolic and Oral Communications

John Elliott and Eric Atwell

Centre for Computer Analysis of Language And Speech,
School of Computer Studies, University of Leeds, Leeds, United Kingdom

This paper describes algorithms and software to characterise and detect generic intelligent language-like features in an input signal, using Natural Language Learning techniques: looking for characteristic statistical "language-signatures" in a test corpus. We assume "language" is not restricted to humans, but that language-like features are to be detected in signals from other intelligent species, including birds, dolphins, and ultimately Extra-Terrestrials. As a first step towards such species-independent language-detection, we present a suite of programs to analyse digital representations of a range of symbolic and oral data, and use the results to extrapolate whether or not there are language-like structures which distinguish this data from other sources, such as music, images, and white noise. We assume that generic species-independent communication can be detected by concentrating on localised patterns and rhythms, identifying segments at the level of characters, words and phrases, without necessarily having to "understand" the content.

As inter-species comparators with human language, we would have preferred alien messages found by the Search for Extra-Terrestrial Intelligence, but none were available. Instead, recordings from birds and dolphins were used, as psychologists and linguists indicate that birds and dolphins share some human developmental and social imperatives for communication. They learn language beyond innate imperative cries they are born with, and their communications demonstrate individual signatures and variations between family groups, reflecting social structures.

We assume that a language-like signal will be encoded either symbolically, i.e. some kind of character-stream; or else as a digitised audio signal, i.e. some kind of sound recording. Our language-detection algorithm for symbolic input uses a number of statistical clues: data-compression ratio, "chunking" to find character-bit length and boundaries, and matching against a Zipfian type-token distribution for "letters" and "words". To detect language-like characteristics in a digitised audio signal, our language-detection algorithm looks for clues in the wave-form, analysed into Significant Activity Sessions (SASs). SASs are "packets" of language in the audio signal, characterised by Amplitude and Duration. The type-token distribution of SAS Amplitudes for human speech is similar to that of dolphin sounds and birdsong, yet clearly different to that of music or white noise. Similarly, the graphs of SAS Durations over time for data from humans, dolphins and birdsong show familial similarity, yet clearly differ from Duration-distributions for music and white noise.

We do not claim extensive (let alone exhaustive) empirical evidence that our language-detection clues are "correct"; the only real test will come when the Search for Extra-Terrestrial Intelligence finds true alien signals. If and when true SETI signals are found, the first step to interpretation is to identify the language-like features, using techniques like the above. Our next research goal is to apply Natural Language Learning techniques to identification of "higher-level" grammatical and semantic structure in a linguistic signal.