IAA-01-IAA.9.2.07

FIRST STAGE IDENTIFICATION OF SYNTACTIC ELEMENTS IN AN EXTRA-TERRESTRIAL SIGNAL

John Elliott, jre@comp.leeds.ac.uk
Eric Atwell, eric@comp.leeds.ac.uk, Bill Whyte, billw@comp.leeds.ac.uk
Centre for Computer Analysis of Language and Speech,
School of Computer Studies, University of Leeds, Leeds, Yorkshire, LS2 9JT England

Having detected a signal from an Extra-Terrestrial source, which satisfies criteria indicating language-like structures at a physical level [Elliott, Atwell '99], second stage analysis is required to begin the process of identifying internal grammar components which constitute the basic building blocks of the symbol system.

Unlike traditional natural language processing, a solution cannot be assisted using vast amounts of training data with well-documented 'legal' syntax and semantic interpretation. Using computational linguistic universals derived from analysing a representative sample set of the human chorus, algorithms developed are designed to work unsupervised and without in-built prior knowledge, for the filtration of inter-galactic Objets Trouv s and the decoding of an unknown signal's grammar structure.

With the use of embedded clauses and phrases, we are able to represent an expression or description, however complex, as a single component of another description. This allows us to build up complex structures far beyond our otherwise restrictive cognitive capabilities. It is this universal hierarchical structure, together with the essential ontological requirements for describing the world around us (external), and the relationships between grammatical structures (internal) evident in all human languages, and necessary for any advanced communicator, that constitute the next phase in the signals analysis.

Given this, the first step to interpretation is to identify these language-like features, detecting where word chunks and phrase-like boundaries occur. It is from these basic syntactic units that the analysis of behavioural trends and inter-relationships amongst terminals and non-terminals alike begins to unlock the encoded internal grammatical structure; clustering into syntacto-semantic classes and indicating candidate parts of speech.

Therefore, at this stage we are endeavouring to identify features of language surface structure, which are universal - or at least quasi-universal - to the world's major language families, irrespective of their adopted scripts or lexical encoding strategies. It is by using these features, that algorithms developed from such information will enable us to extract core syntactic elements without the necessity of a primer or universal Rossetta stone and making as few assumptions as possible.

In addition to the search for such decoding strategies of language scripts and thereby the understanding of what language structure actually is, is the analysis of audio signals for their structural features to facilitate the discrimination of language-like signals from non-language phenomena. To further enable rigorous analysis of this problem we have continued to look at a variety of sound/signal samples, which also include the 'sounds of space' tape compiled by Dr Cullers. To date our results remain encouraging and we are continuing to positively discriminate language - whether distorted or subject to interference - from 'other' natural phenomena.

This paper describes the algorithms and software developed for these purposes, including a visualisation tool to facilitate the examination of annotation-combination space topology and approach vectors.

Reference:

Elliott, J. and Atwell, E. (1999), "Language in signals: the detection of generic species-independent intelligent language features in symbolic and oral communications." In Proceedings of the 50th International Astronautical Congress. Amsterdam, paper IAA-99-IAA.9.1.08.