

Takac, M., Benuskova, L., Knott, A.: Sure as Eggs is Eggs: Low Entropy as Predictor of Idiomaticity in Language Production by a Simple Recurrent Network, *The 17th Australian Language and Speech Conference (ALAS'09)*, Sydney, 3-4 Dec 2009

Sure as eggs is eggs: Low entropy as predictor of idiomaticity in language production by a simple recurrent network

Idioms are continuous ('kick the bucket') or discontinuous ('take NP to task') fixed multi-word expressions whose meaning is defined holistically, rather than through the meanings of individual words. In this paper we present a neural network model of sentence generation, which is able to produce both continuous and discontinuous idioms within regular compositional sentences. The model is a simple recurrent network, extended to include a semantic episode representation as an extra input. It is trained on (artificially generated) pairs of episode representations and word sequences. The network learns to rely on a mixture of semantic and syntactic representations to generate compositional (non-idiomatic) language, and on its own context representations to generate idiomatic language. To confirm the hypothesis that idiomatic phrases can be completed based on syntactic properties only, we examined the performance of a network with no semantic inputs, trained for next-word prediction task on both the artificial language and a real corpus. Interpreting the network's output as next-word probability distribution and the entropy of the distribution as an inverse degree of self-confidence of the network's prediction, idioms had the special status of very low entropy within the generated language.