Crossmodal Entropy Transfer

Bo Pedersen (bop@cornell.edu)
Department of Psychology, Cornell University
Uris Hall, Ithaca, NY 14853 USA

Empirical models of language acquisition traditionally focus on learning the syntax of "clean" text corpora, without the rich environmental information actually available to children. Under these minimal conditions, this possibility for learning faces a body of work in theoretical computer science more or less rendering it impossible in idealized cases (context dependent grammars) and thus lends support to more nativist theories of language (e.g., Gold 1967). However, the work of Christiansen and Dale (2001) suggests that if some of the contextual information is put back into the equation it might become solvable again. Simulations with simple-recurrent networks indicate that the integration of distributional, phonological and prosodic cues can promote better, faster and more uniform learning. Recent studies by Howel, Jankowicz and Becker (2005) suggest that the inclusion of sensorimotor cues like hard, soft, in-kitchen, man-made, etc. can improve syntax learning in a similar setup, and Pedersen (2006) suggest that the same might be accomplished with image features extracted from a set of real world scenes collected with a head mounted webcam.

In this paper we demonstrate a method that can show the existence and strength of such cues beyond the model grammars and model cues mentioned above. If we could mount a video camera on a child's head and record all the visual and auditory stimuli the child is exposed to during a day we would have an excellent corpus regarding those natural visual cues. Unfortunately such data do not exist – but we do have access to a corpus of co-occurring visual and auditory stimuli that constitutes a great deal of modern reality, namely television. TV shows and movies can be recorded and subtitles can be extracted or parsed. While we wait for the ultimate language-context corpus to emerge, TV may serve as a test bed for some of these hypotheses.

In this study, we made use of a short TV sequence for this purpose. 5000 features were picked by random from a 4 minute cooking video (Martha Stewart's "Roast Chicken"). Each feature was a 10x10 pixel square picked from a random position in a random frame. A time series was then created from the errors of fitting this feature optimally into each frame of the video, thus giving us 5000 time series. The idea is that some of the features will represent something meaningful, for example the edge of a lemon, so that the error of this feature is low when there is a lemon in the picture and high otherwise.

This video was transcribed and the 256 most frequent words encoded into 16x16 doublets so that the transcript is represented by a time series of numbers between 1 and 16 (to reduce the granularity and complexity of the transfer entropy computations). We then have two time series, a text corpus and a video encoding that can be tested for couplings using Transfer Entropy (Schreiber, 2000). We tested a series of alignments between the text and all of the video features, close to their co-occurrence in time, and a t-test performed on the points in the two series before and after the 0 alignment showed a significant difference at the 0.001 level between the before and after points.

Note that no assumptions about language or vision have been made. For language, words are represented by numbers only and for vision we have some error rates of some random example based features and still we can see a coupling between the two, even though the nature of the two time series are substantially different. So not only do we have statistical regularities in our visual world and in language as documented by many scholars, but we also have statistical regularities that bind these two modalities together. In the way that these regularities facilitate learning in language and vision, we speculate that couplings like these can facilitate the learning of both and in particular the relation between them, so that when we have just a speaker and no images we have previous couplings than can help us create imagery and when we have images only, words will come to mind.

In the future we will process many more videos like this. But although this analysis is based on one video only it demonstrates a technique that can potentially be used to create existence proofs of couplings between modalities for real empirical data and even quantify these couplings.

Christiansen, M.H. & Dale, R.A.C. (2001). Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp. 220-225). Mahwah, NJ: Lawrence Erlbaum

Gold, E. (1967) *Language identification in the limit*, Information and Control, vol.10, 447-474

Howell, S. R. & Jankowicz, D. & Becker, S. (2005). A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning. *Journal of Memory and Language* (pp. 258-276). Orlando, FL: Elsevier.

Pedersen, B. (2006) *Visual Cues in Connectionist Models of Language Acquisition*. In the 28th Annual Meeting of the Cognitive Science Society, CogSci2006 Proceedings, 2006.

Schreiber, T. (2000) *Measuring information transfer*, Physical Review Letters, 85(2), July 2000, 461-464