Information Theory

Robin Clark

Final version

INFORMATION THEORY rests on the fundamental observation that information and uncertainty are related (Shannon and Weaver, 1949). Intuitively, a code can be used to send information from one agent (the transmitter) to another (the receiver) or a channel just in case the receiver cannot completely anticipate which message the transmitter will send. A "language" that consisted of only one sentence could not be a useful instrument of communication precisely because there could be neither a real choice (on the part of the transmitter) nor any real uncertainty (on the part of the receiver) about which message could be sent.

Entropy is a measure of the uncertainty in a communication system. Given that uncertainty and information can be identified, we can say that a measure of the uncertainty in a system is also a measure of its information content. Suppose that a communication system provides n distinct symbols and that p_i is the probability that the i^{th} symbol occurs; then the entropy, H, is given by:

$$H = -\sum_i p_i \log p_i$$

H can vary between 0 and log(n); the closer H is to log(n), the greater the uncertainty of the system. In other words, if H were log(n), this would imply that the choice of the next symbol is random (with uniform probability). Clearly, this does not hold for natural languages where relationships like agreement, case-marking and selectional restrictions serve to limit the speaker's choice about what form to pick at any given point.

Early information theoretic approaches natural language assumed that the processes that limit the speaker's choice were local. That is, natural languages were hypothesized to be *stationary* in the sense that statistical

effects on choice would be restricted to a relatively small window that could be modeled by digrams (pairs of words) or trigrams (triples of words) → *Hidden Markov Models*. Chomsky (1956), however, argued persuasively that natural languages are not stationary; indeed, languages could contain unbounded nested dependencies, as witnessed by the relationship between pairs like *both...and...* and *either...or...*.

Despite this obvious limitation, the conclusion that information theoretic techniques are irrelevant to the study of language is unwarranted. Information theoretic considerations have been used to disambiguate syntactic structures, learn attachment preferences for modifiers, disambiguate word senses, sort words into grammatical categories and into classes based on lexical semantics (see Charniak 1993 for a useful review of applications to → *Computational Linguistics*) and play a dominant role in → *Automatic Speech Recognition* and → *Optical Character Recognition*. To take one example, encountering a verb tends to reduce the uncertainty of the surrounding environment since verbs place strong syntactico-semantic restrictions on their environment. In verb-second languages, however, the verb occurs in a special position where it places no constraints on the items that immediately precede it. Thus, entropy around the verb can be used to distinguish verb-second from non-verb-second languages (Brill and Kapur 1993).

Despite its obvious interest, information theory has yet to be integrated into mainstream theoretical linguistics. One possible direction for doing so is suggested by *Kolmogorov complexity*, a precise mathematical realization of Occam's razor (Cover & Thomas, 1991; Li and Vitányi, 1993), with far reaching consequences for linguistic typology as well as the theory of learnability. It should be noted, however, that Kolmogorov complexity can be determined only within an additive constant, which limits the applicability of the concept to cases where complexity grows without bounds so that the constant can be neglected.

References Brill, E., and Kapur, S. An information theoretic solution to parameter setting. IRCS Technical Report, 93–07, University of Pennsylvania, 1993. Institute for Research in Cognitive Science.

Charniak, E. *Statistical Language Learning*. The MIT Press, Cambridge, MA, 1993.

Chomsky, N. Three models for the description of language. *IRE Transactions on Information Theory 2* (1956), 113–124.

Cover, T. M., and Thomas, J. A. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, 1991.

Li, M., and Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications.* New York, NY, 1993.

Shannon, C. E., and Weaver, W. W. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, IL, 1949.

Index Information Theory, entropy, Kolmogorov complexity