A Semantic Markov Field Model of Text Recall

Richard M. Golden
Cognition & Neuroscience Program, GR41
School of Human Development
University of Texas at Dallas
Richardson, TX 75083-0688
golden@utdallas.edu

Abstract

A probabilistic model of text recall is proposed which assigns a probability mass to a given recall protocol. Knowledge analyses of semantic relationships among events identified in the text are used to specify the architecture of the probability model. Twelve subjects (the training data group) were then asked to recall twelve texts from memory. The recall protocols generated by the twelve subjects were then used to estimate the strengths of the semantic relationships in the probabilistic model. The Gibbs Sampler algorithm (a connectionist-like algorithm) was then used to sample from the probabilistic model in order to generate synthesized recall protocols. These synthesized recall protocols were then compared with the original set of recall data and recall data collected from an additional group of twelve human subjects (the test data group).

Markov Field Probability Model

In this section, a Markov field probability model which assigns a probability mass to a given recall protocol (i.e., an ordered sequence of complex text propositions or equivalently *text features*) is now explicitly defined. The specific formulas described in this section are discussed in detail and derived elsewhere. The semantic associative links (e.g., causal links) identified by a text knowledge analysis are first used to specify the free parameters of a special d-dimensional matrix which is called the $\mathbf{W}$ matrix. The ijth element of $\mathbf{W}$ indicates the degree to which the activation of the jth text feature, y_j, in the *working memory buffer*, $\mathbf{y} = [y_1, \ldots, y_d]$, influences the probability that the ith text feature will be recalled. Note that only a subset of elements of $\mathbf{W}$ are estimated using quasi-maximum likelihood estimation (i.e., the a priori designated semantic and episodic knowledge links) while the remaining elements of $\mathbf{W}$ are constrained to be equal to zero.

Let $\mathbf{f}_i$ be a d-dimensional vector with a one in position i and zeros in all remaining $d-1$ positions. The vector $\mathbf{f}_i$ identifies the ith text feature in the text. Let the notation $\mathbf{x}(t) = \mathbf{f}_j$ indicate that the tth item in a human subject's recall protocol was text feature $\mathbf{f}_j$. The working memory buffer of the human subject is assumed to be updated according to the formula $\mathbf{y}(t+1) = \mathbf{x}(t) + \mu\mathbf{y}(t)$ where the empirically determined $\mu \in [0, 1)$ specifies the working memory node activation *decay rate*.

Let $\mathbf{y}(t) = [y_1(t), \ldots, y_d(t)]$ model the working memory buffer of the human subject after t items have been recalled. Define the *local potential function* $V_{s,i}$ such that:

$$V_{s,i} = \mathbf{f}_i^T[\mathbf{W}\mathbf{y}(s) + \sum_{t=s+1}^{M} [\mathbf{W}\mu^{t-1-s}]^T\mathbf{x}(t)]. \quad (1)$$

Then the conditional probability that the sth item in a recall protocol is the ith text feature recalled given knowledge of all other items in the recall protocol is given by the formula:

$$p_{s,i} = \frac{exp[V_{s,i}]}{\sum_{k=1}^{d} exp[V_{s,k}]}. \quad (2)$$

Probability Model Evaluation

Quasi-maximum likelihood estimates of the strengths of semantic and episodic knowledge links in the $\mathbf{W}$ matrix were computed from recall protocol data collected from twelve human subjects (*training group*) for twelve short 12-15 sentence texts. The Geman and Geman Gibbs Sampler algorithm was then used to sample from the joint probability distribution of recall protocols. An additional (*test group*) of twelve human subjects also recalled the twelve texts from memory. The recall protocols generated from the two groups of human subjects and the recall protocols generated from the Gibbs Sampler algorithm were then compared.

Statement recall probabilities computed from Gibbs Sampler generated recall protocols were *quantitatively* very similar to statement recall probabilities computed from actual human recall protocols. Both human subject recall protocols and Gibbs Sampler generated recall protocols exhibited the well-known finding that statements with more causal connections to other statements in a text are more likely to be included in a recall protocol.

The synthesized recall protocol data may also be directly compared with the human recall protocol data in order to make predictions about the *explicit order* in which items are recalled from memory. For example, the sixth synthesized recall protocol for the *miser* text is the text feature sequence $\mathbf{f}_3, \mathbf{f}_7, \mathbf{f}_{14}, \mathbf{f}_{16}, \mathbf{f}_{17}$ which corresponds to the sequence of English statements:

> The miser buried the gold in the ground ($\mathbf{f}_3$). The servant stole the gold ($\mathbf{f}_7$). The neighbor said, "Go and take a stone, and bury it in the hole" ($\mathbf{f}_{14}$). The neighbor said, "The stone will be as useful to you as the gold"($\mathbf{f}_{16}$). The neighbor said, "When you had the gold, you never used it"($\mathbf{f}_{17}$).