Technical Report: Annotated Semantic Markov Utterance Random Fields for Information Extraction

Richard M. Golden, *Senior Member, IEEE*

***Abstract*—This technical report presents the mathematical foundations of the ASMURF (Annotated Semantic Markov Utterance Random Field) methodology for information extraction. The ASMURF semantic annotation system is presented, and key theorems establishing the computational adequacy of the methodology are presented.**

***Index Terms*—Information Extraction, Natural Language Understanding, Hidden Markov Models, Markov Random Fields, Semantic Annotation, Gibbs Distribution**

I. Introduction

INFORMATION extraction (IE) technology has received increasing amounts of attention over the past several years. In contrast to information retrieval (IR) technology which is concerned with the problem of automated document retrieval, IE technology is concerned with the problem of automated information retrieval. IE technology also differs fundamentally from the problem of solving the full-blown natural language understanding problem. The general natural language understanding problem is concerned with developing computer systems which have a ”deep” understanding of a text. In contrast, IE technology is not concerned with the problem of trying to understand all information conveyed in a text. IE technology is concerned with simply attaining a ”partial understanding” of a text for the purpose of extracting specific information. IE technology can be applied in a range of situations. Examples of IE technology include interpreting natural language queries for web searches, improving the performance of speech recognition systems, developing natural language interfaces for medical and travel applications, and the analysis of student essay data.

Hidden Markov Models (HMMs) have been used for several decades in areas such as speech perception (e.g., Baum et al., 1970; Annals of Mathematical Statistics, 41, 164-171; Rabiner, 1989, Proc.IEEE, 77, 1989, 257-285). However only recently have HMM models been applied in the context of IE applications. In the past five years, HMM Information Extraction (IE) methods have been used for topic detection and tracking (Yamron et al., 1998), dialog act modeling (Stolcke et al., 1998), search scientific documents for gene names and gene locations (Leek, 1997), extracting information from document headers (Seymore, McCallum, and Rosenfeld, 1999; Freitag and McCallum, 1999), student essay analysis (Burstein, Marcu, and Knight, 2000), and recall data (Durbin et al., 2000), and document summarization (Schlesinger et al., 2000). Moreover, the formulation of previous HMM methods have been based upon Markov Chain modeling methods in contrast to the more general Markov Random Field (MRF) methodology. Hidden Markov Random Fields have been extensively applied in image processing applications but this technology has not yet been applied to solve real-world computational linguistics engineering problems.

The specific IE problem considered in this paper is concerned with situations where: (1) large amounts of semantically annotated training data are not available, (2) the documents to be automatically semantically annotated consist of highly ungrammatical sentence fragments and misspelled words, (3) a detailed semantic annotation of the data is desired, and (4) considerable detailed domain knowledge regarding document content is available in advance. An important application area associated with this type of IE problem is the extraction of information from surveys consisting of open-response items and essay exams.

The author is with the School of Behavioral and Brain Sciences (GR4.1), University of Texas at Dallas (Box 830688), Richardson, TX, 75083-0688, USA. Email: golden@utdallas.edu

This research was funded by the NSF Information Technology Research Learning and Education Program Award 0113369.

Golden, R. M. (2006). Annotated Semantic Markov Utterance Random Fields for Information Extraction. Technical Report: Technical Report: Annotated Semantic Markov Utterance Random Fields for Information Extraction, School of Behavioral and Brain Sciences (GR4.1), University of Texas at Dallas, Richardson, TX 75083-0688

This paper is organized in the following manner. First, the proposed semantic annotation system. Second, the probabilistic knowledge representation assumptions are formulated within a novel Hidden MRF framework. Third, the use of the novel MRF framework for inferring the Maximum A Posteriori (MAP) semantic annotation is described. Fourth, efficient methods for Maximum A Posteriori parameter estimation using the MRF framework are formulated as well. The computational performance of the resulting system is then evaluated with respect to the experimental data set.

A. Semantic Annotation System

The semantic annotation system which is used involves three basic forms of concepts: (1) ”word-concepts”, (2) ”atomic propositions”, and (3) ”molecular propositions”. The objective of the semantic annotation process is formally defined in terms of molecular propositions. Specifically, the goal of the semantic annotation process is to map a word sequence (e.g., free response data generated by a student in response to an essay question) into an ordered sequence of molecular propositions. This ultimate objective is achieved by first semantically annotating key words with their unambiguous semantic ”word-concept” interpretations, and then annotating sequences of word-concepts with their unambiguous ”atomic proposition” interpretations. It is assumed that some words (particularly words which have especially ambiguous and subtle grammatical functions) will be ignored. Such words are assigned to a ”skip word list”. The semantic annotation system is always problem-specific which means that an entirely new semantic annotation system must be developed for each application.

In a practical application, a sample of training data is used to identify an initial set of molecular propositions. These molecular propositions, in turn, are used to specify an initial set of atomic propositions which, in turn, are used to specify an initial set of word-concepts. This initial concept dictionary is then embedded within an interactive user-friendly graphical user-interface intended to solicit refinements of the semantic annotation scheme from human semantic annotators or ”coders”. An example of the user-interface is shown in Figure 1. Specific details of the semantic annotation system embodied in the software are provided in the remainder of this section.

1) Word-Concepts: A *w*ord-concept is a label for a set of words which are considered to be approximately semantically equivalent for a particular semantic annotation application. For example, in one application, the words $canteen$, $cafeteria$, and $restaurant$ might be considered to be semantically equivalent and assigned the common word-concept **CAFETERIA**. In other applications, however, making semantic distinctions between $cafeteria$ and $restaurant$ might be essential for an appropriate semantic annotation. Word-concept categories are useful for clarifying semantic distinctions among word-concepts. The word-concept categories used in the current semantic annotation system are: **ACTION-MODIFIER**, **AGENT**, **ATTRIBUTE**, **COMMUNICATIVE-ACTION**, **EXPERIENCER**, **INSTRUMENT**, **LOCATION**, **MENTAL-ACTION**, **OBJECT**, and **PHYSICAL-ACTION**.

Assume there are d_c word-concepts to be represented in the semantic annotation system and that the mth word-concept is denoted by the mth column, $\mathbf{c}^{(m)}$, of the *w*ord-concept dictionary which is a d_c-dimensional identity matrix, $m = 1, \ldots, d_c$. Similarly, assume there are d_w words to be represented in the semantic annotation system and that the mth word is denoted by the mth column, $\mathbf{w}^{(m)}$, of the *w*ord dictionary which is a d_w-dimensional identity matrix, $m = 1, \ldots, d_w$.

A $word - concept\ random\ vector$ is a discrete d_w-dimensional random vector which takes on the value of the mth column of the word-dictionary with a strictly positive probability for $m = 1, \ldots, d_w$. The notation $\mathbf{c}_{t,i,j}$ denotes the jth word-concept within the ith atomic proposition located within the tth molecular proposition.

2) Atomic Propositions: An $atomic\ proposition$ is a label for a set whose elements are approximately semantically equivalent sequences of one or more word-concepts for a particular semantic annotation application. Just as semantic annotation decisions regarding equivalence classes of words associated with word-concepts must be made, semantic annotation decisions regarding equivalence classes of sequences of word-concepts must be made as well. For example, the word-concept sequences:

$\{$ **AGENT:ESPERANZA, ACTION:EAT , OBJECT:LUNCH** $\}$

and

{ **AGENT:FEMALE, ACTION:EAT, OBJECT:LUNCH** }

might be considered to be members of the same equivalence class if it is known that the discourse context is constrained such that the only female person who eats lunch in the discourse context is in fact ”Esperanza”.

In other applications, where multiple agents might be catching the ball, however, these two word-concept sequences would not be considered to be semantically equivalent. Atomic propositions are defined in the current semantic annotation system as propositions which either: (1) refer to exactly one action word-concept (i.e., mental action, communicative action, or physical action), or (2) describe a state of the environment using attribute word-concepts. Assume there are d_a word-concepts to be represented in the semantic annotation system and that the mth word-concept is denoted by the mth column, $\mathbf{a}^{(m)}$, of the *a*tomic proposition dictionary which is a d_a-dimensional identity matrix, $m = 1, \ldots, d_a$.

An $atomic\ proposition\ random\ vector$ is a discrete d_a-dimensional random vector which takes on the value of the mth column of the atomic proposition dictionary with a strictly positive probability for $m = 1, \ldots, d_a$. The notation $\mathbf{a}_{t,i}$ denotes the ith atomic proposition in a sequence of atomic propositions which expresses a representative molecular proposition.

3) Molecular Propositions: A $molecular\ proposition$ is a label for a set whose elements are approximately semantically equivalent sequences consisting of one or more atomic propositions for a particular semantic annotation application. For example, in one application, the sequence of two atomic propositions: **REQUEST(AGENT:ESPERANZA)**, **EAT(AGENT:ESPERANZA,OBJ:FOOD)** might be considered to be semantically equivalent to the atomic proposition sequence: **EAT(AGENT:ESPERANZA,OBJ:FOOD) REQUEST(AGENT:ESPERANZA)**,

Assume there are d_f molecular propositions to be represented in the semantic annotation system and that the mth molecular proposition is denoted by the mth column, $\mathbf{f}^{(m)}$, of the *m*olecular proposition dictionary which is a d_f-dimensional identity matrix, $m = 1, \ldots, d_f$.

A $molecular\ proposition\ random\ vector$ is a discrete d_f-dimensional random vector which takes on the value of the mth column of the molecular proposition dictionary with a strictly positive probability for $m = 1, \ldots, d_f$. The notation $\mathbf{f}_t$ denotes the tth molecular proposition in a sequence of molecular propositions generated by a participant (or group of participants) within the essay question free response paradigm.

II. Markov Random Field Formulation

Referring to Figure 1, consider an example where a sequence of words mentioned by the student is:

$$esperanza\ wanted\ to\ eat\ in\ the\ canteen$$

and assume it is known that this word sequence corresponds the third molecular proposition mentioned by student 1 in Table 2. In this example, the words to, in, and the are on the *skip word list* and thus will be ignored. Thus, the remaining words in the word sequence $esperanza\ wanted\ eat\ canteen$ must be assigned word-concepts. Also assume that it is known that the word subsequences $esperanza\ wanted$ and $eat\ canteen$ are associated with two distinct atomic propositions. The semantic annotation problem is to assign word-concepts to the 4 words which are not on the skip list, assign atomic propositions to the 2 subsequences of words, and assign a molecular proposition to the entire word sequence. The system's performance will be evaluated primarily with respect to the appropriateness of the system's choice for the molecular proposition.

The molecular proposition random vector $\tilde{\mathbf{f}}_3$ has a probability distribution which is functionally dependent upon the previously assigned values to molecular propositions $\tilde{\mathbf{f}}_2$ and $\tilde{\mathbf{f}}_1$ as well as the sequence of atomic propositions $\tilde{\mathbf{a}}_{3,1}$ and $\tilde{\mathbf{a}}_{3,2}$. The probability distribution of the atomic proposition random vector $\tilde{\mathbf{a}}_{3,2}$ is functionally dependent upon atomic proposition $\tilde{\mathbf{a}}_{3,1}$, molecular proposition $\tilde{\mathbf{f}}_3$, and the sequence of word-concepts $\tilde{\mathbf{c}}_{3,2,1}$ followed by $\tilde{\mathbf{c}}_{3,2,2}$. The probability distribution of word-concept random vector $\tilde{\mathbf{c}}_{3,1,2}$ is functionally dependent upon the word random vector $\tilde{\mathbf{w}}_{3,1,2}$.

The probabilistic modeling assumptions of the proposed solution to the semantic annotation problem are naturally formulated within a Markov Random Field framework.

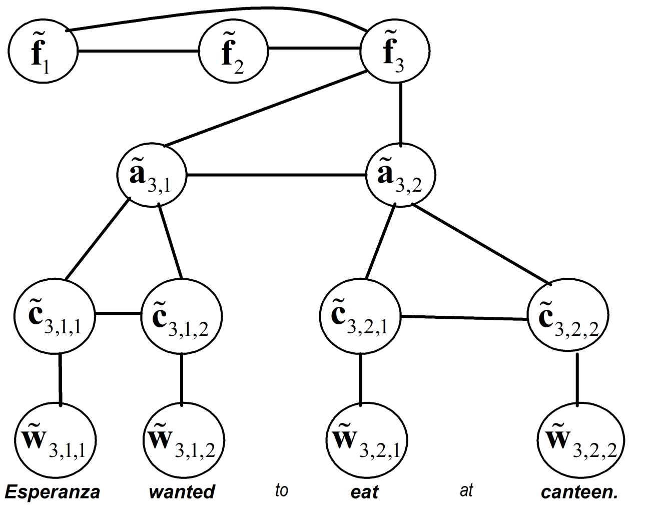
$\tilde{\mathbf{f}}_1$
$\tilde{\mathbf{f}}_2$
$\tilde{\mathbf{f}}_3$
$\tilde{\mathbf{a}}_{3,1}$
$\tilde{\mathbf{a}}_{3,2}$
$\tilde{\mathbf{c}}_{3,1,1}$
$\tilde{\mathbf{c}}_{3,1,2}$
$\tilde{\mathbf{c}}_{3,2,1}$
$\tilde{\mathbf{c}}_{3,2,2}$
$\tilde{\mathbf{w}}_{3,1,1}$
$\tilde{\mathbf{w}}_{3,1,2}$
$\tilde{\mathbf{w}}_{3,2,1}$
$\tilde{\mathbf{w}}_{3,2,2}$
Esperanza
wanted
to
eat
at
canteen.

Fig. 1. MRF representation for analyzing sentence: ”Esperanza wanted to eat at the canteen.”

A. Markov Random Fields

An $undirected\ graph$ is defined as a set $G = (S, E)$ consisting of: (1) a set of $nodes$ S, and (2) a set of $edges$ $E \subseteq S \times S$. A MRF (Markov Random Field) is a pair $(\tilde{\mathbf{X}}, G)$ where $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_d\}$) and G is an undirected graph called the MRF's *n*eighborhood graph. MRF's are assumed to satisfy the *p*ositivity condition that the probability of every possible realization of the field is strictly positive (i.e., $p(\tilde{\mathbf{X}}) > 0$). The positivity condition ensures that all conditional probability distributions defined with respect to the MRF exist.

Given a neighborhood graph $G = (S, E)$ for a MRF, the $neighborhood$ for the ith random variable, $\tilde{\mathbf{x}}_i$, in the MRF is defined as the set

$$\mathcal{N}_i = \{\tilde{\mathbf{x}}_j \in \tilde{\mathbf{X}} : (\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \in E\}.$$

Conditional independence assumptions in the MRF are specified by the choice of E since it is assumed that:

$$p(\tilde{\mathbf{x}}_i | \mathcal{N}_i) = p(\tilde{\mathbf{x}}_i | \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{i-1}, \tilde{\mathbf{x}}_{i+1}, \ldots, \tilde{\mathbf{x}}_d).$$

B. Annotated Semantic Markov Utterance Random Field (ASMURF) Problem Representation

1) Segmentation Strategy Graph:

Definition 1: Clique Let G be a finite undirected graph. A *clique* of G is defined as a non-empty set C such that either: (1) C contains exactly one node of G, or (2) every pair of nodes in C is an edge of G.

Definition 2: ASMURF A *Annotated Semantic Markov Utterance Random Field* or (ASMURF) is a MRF comprised of the following random vectors.

- A random vector, $\tilde{\mathbf{f}}$, comprised of d molecular proposition random subvectors $\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_d$.
- A random vector, $\tilde{\mathbf{a}}$, comprised of d random subvectors such that the ith random subvector, $\tilde{\mathbf{a}}_i$, consists of the d_i atomic proposition random subvectors $\tilde{\mathbf{a}}_{i,1}, \ldots, \tilde{\mathbf{a}}_{i,d_i}$ associated with $\tilde{\mathbf{f}}_i$, $i = 1, \ldots, d$.
- A random vector $\tilde{\mathbf{c}}$, comprised of a finite set of word-concept random subvectors $\tilde{\mathbf{c}}_{i,j,1}, \ldots, \tilde{\mathbf{c}}_{i,j,d_{i,j}}$ associated with each $\tilde{\mathbf{a}}_{i,j}$, where $i = 1, \ldots, d$ and $j = 1, \ldots, d_i$.
- A random vector $\tilde{\mathbf{w}}$, comprised of a finite set of word random subvectors $\tilde{\mathbf{w}}_{i,j,k}$ for each word-concept random vector $\tilde{\mathbf{c}}_{i,j,k}$, $i = 1, \ldots, d$, $j = 1, \ldots, d_i$, and $k = 1, \ldots, d_{i,j,k}$.

The graph G associated with a ASMURF is called the *s*egmentation strategy for the ASMURF.

C. ASMURF Specifications

In this section, the global and local specifications of a ASMURF are provided.

1) ASMURF Joint Distribution: The joint probability mass function, $p : S \to [0, 1]$ of the random variables in a SA-MRF whose parameter vector β is given by:

$$p(\mathbf{f}, \mathbf{a}, \mathbf{c}, \mathbf{w} | \beta) = Z^{-1} exp[-V(\mathbf{f}, \mathbf{a}, \mathbf{c}, \mathbf{w})] \qquad (1)$$

where

$$Z = \sum_{[\mathbf{f}, \mathbf{a}, \mathbf{c}, \mathbf{w}] \in S} exp[-V(\mathbf{f}, \mathbf{a}, \mathbf{c}, \mathbf{w})]$$

and where

$$V(\mathbf{f}, \mathbf{a}, \mathbf{c}, \mathbf{w}) = \beta_f (1/d) \sum_{t=1}^{d} V_f(\mathbf{f}_{t-2}, \mathbf{f}_{t-1}, \mathbf{f}_t) +$$

$$\sum_{t=1}^{d} \sum_{i=1}^{d_t} V_a(\mathbf{a}_{t,i-2}, \mathbf{a}_{t,i-1}, \mathbf{a}_{t,i}, \mathbf{f}_t) +$$

$$\sum_{t=1}^{d} \sum_{j=1}^{d_t} \sum_{k=1}^{d_{t,j}} V_c(\mathbf{c}_{t,j,k-2}, \mathbf{c}_{t,j,k-1}, \mathbf{c}_{t,j,k}, \mathbf{a}_{t,j}) +$$

$$\sum_{t=1}^{d} \sum_{j=1}^{d_t} \sum_{k=1}^{d_{t,j}} \sum_{k=1}^{d_{t,j}} V_w(\mathbf{c}_{t,j,k}, \mathbf{w}_{t,j,k}).$$

The functions V_f, V_a, V_c, and V_w are local potential functions of the ASMURF and are defined in the following sections. The vectors $\mathbf{f}_0$, $\mathbf{f}_{-1}$, $\mathbf{a}_{t,0}$, $\mathbf{a}_{t,-1}$, are defined to be vectors of zeros. It will also be convenient to define $\mathbf{f}_{d+1}$ and $\mathbf{f}_{d+2}$ equal to vectors of zeros.

2) Molecular Potential Function: The *molecular proposition potential function* $V_f : D_f \times D_f \times D_f \to \mathcal{R}$ for the MRF is defined by the formula:

$$V_f(\mathbf{f}_{t-2}, \mathbf{f}_{t-1}, \mathbf{f}_t) = \mathbf{f}_t^T[\mathbf{b}_0^f + \mathbf{B}_{-1}^f \mathbf{f}_{t-1} +$$

$$\mathbf{B}_{-2}^f \mathbf{f}_{t-2} + \mathbf{B}_{-1,-2}^f(\mathbf{f}_{t-1} \otimes \mathbf{f}_{t-2})] \qquad (2)$$

where the matrices $\mathbf{B}_{-1}^f, \mathbf{B}_{-2}^f, \mathbf{B}_{-1,-2}^f$ and vector $\mathbf{b}_0^f$ are constants which specify the functional form of V_f. The notation $\otimes$ denotes the Kronecker tensor product which is defined such that: $\mathbf{A}\otimes\mathbf{B}$ is a matrix of submatrices where the ijth submatrix is defined by $a_{ij}\mathbf{B}$.

The *local molecular probability mass function* $p_f : D_f \to [0, 1]$ for the ASMURF is defined such that:

$$p_f(\mathbf{f}_t | \mathbf{f}_{t+1}, \mathbf{f}_{t+2}, \mathbf{f}_{t-1}, \mathbf{f}_{t-2}, \mathbf{a}_{t,1}, \ldots, \mathbf{a}_{t,d_t}) =$$

$$\frac{exp[\phi_{\mathbf{f}_t}]}{\sum_{\mathbf{f}\in D_f} exp[\phi_{\mathbf{f}}]}$$

where

$$\phi_{\mathbf{f}_t} = V_f(\mathbf{f}_{t-2}, \mathbf{f}_{t-1}, \mathbf{f}_t) + V_f(\mathbf{f}_{t-1}, \mathbf{f}_t, \mathbf{f}_{t+1}) +$$

$$V_f(\mathbf{f}_t, \mathbf{f}_{t+1}, \mathbf{f}_{t+2}) + \sum_{i=1}^{d_t} V_a(\mathbf{a}_{t,i-2}, \mathbf{a}_{t,i-1}, \mathbf{a}_{t,i}, \mathbf{f}_t).$$

3) Atomic Potential Function: The *atomic proposition potential function* $V_a : D_a \times D_a \times D_a \times D_f \to \mathcal{R}$ for the MRF is defined by the formula:

$$V_a(\mathbf{a}_{t,i-2}, \mathbf{a}_{t,i-1}, \mathbf{a}_{t,i}, \mathbf{f}_t) = \mathbf{a}_{t,i}^T[\mathbf{b}_0^a + \mathbf{B}^a{}_{-1}\mathbf{a}_{t,i-1} +$$

$$\mathbf{B}^a_{-2}\mathbf{a}_{t,i-2} + \mathbf{B}^{a,f}\mathbf{f}_t + \mathbf{B}^a{}_{-1,-2}(\mathbf{a}_{t,i-1} \otimes \mathbf{a}_{t,i-2}) +$$

$$\mathbf{B}_{-1}^{a,f}(\mathbf{a}_{t,i-1} \otimes \mathbf{f}_t) + \mathbf{B}_{-2}^{a,f}(\mathbf{a}_{t,i-2} \otimes \mathbf{f}_t) +$$

$$\mathbf{B}^{a,f}_{-1,-2}(\mathbf{a}_{t,i-1} \otimes \mathbf{a}_{t,i-2} \otimes \mathbf{f}_t)). \qquad (3)$$

where the matrices $\mathbf{b}_0^a$, $\mathbf{B}^a_{-1}$, $\mathbf{B}^a_{-2}$, $\mathbf{B}^{a,f}$, $\mathbf{B}^a_{-1,-2}$, $\mathbf{B}^{a,f}_{-1}$, $\mathbf{B}^{a,f}_{-2}$, and $\mathbf{B}^{a,f}_{-1,-2}$ are constants which specify the functional form of V_a.

The *local atomic probability mass function* $p_a : D_a \to [0, 1]$ for the ASMURF is defined such that:

$$p_a(\mathbf{a}_{t,i} | \mathbf{a}_{t,i-1}, \mathbf{a}_{t,i-2}, \mathbf{a}_{t,i+1}, \mathbf{a}_{t,i+2}, \mathbf{f}_t, \mathbf{c}_{t,i,1}, \ldots, \mathbf{c}_{t,i,d_t}) =$$

$$\frac{exp(\phi_{\mathbf{a}_{t,i}})}{\sum_{\mathbf{a}\in D_a} exp(\phi_{\mathbf{a}})}$$

where

$$\phi_{\mathbf{a}_{t,i}} = V_a(\mathbf{a}_{t,i-2}, \mathbf{a}_{t,i-1}, \mathbf{a}_{t,i}, \mathbf{f}_t) +$$

$$V_a(\mathbf{a}_{t,i-1}, \mathbf{a}_{t,i}, \mathbf{a}_{t,i+1}, \mathbf{f}_t) + V_a(\mathbf{a}_{t,i}, \mathbf{a}_{t,i+1}, \mathbf{a}_{t,i+2}, \mathbf{f}_t) +$$

$$\sum_{j=1}^{d_{t,i}} V_c(\mathbf{c}_{t,i,j-2}, \mathbf{c}_{t,i,j-1}, \mathbf{c}_{t,i,j}, \mathbf{a}_{t,i})$$

4) Word-Concept and Word Potential Functions: The *word-concept potential function* $V_c : D_c \times D_c \times D_c \times D_a \to \mathcal{R}$ for the MRF is defined by the formula:

$$V_c(\mathbf{c}_{t,i,j-2}, \mathbf{c}_{t,i,j-1}, \mathbf{c}_{t,i,j}, \mathbf{a}_{t,i}) = \mathbf{c}_{t,i,j}^T[\mathbf{b}_0^c + \mathbf{B}_{-2}^c \mathbf{c}_{t,i,j-2} +$$

$$\mathbf{B}_{-1}^c \mathbf{c}_{t,i,j-1} + \mathbf{B}^{c,a}\mathbf{a}_{t,i} + \mathbf{B}_{-1,-2}^c(\mathbf{c}_{t,i,j-1} \otimes \mathbf{c}_{t,i,j-2}) +$$

$$\mathbf{B}_{-1}^{c,a}(\mathbf{c}_{t,i,j-1} \otimes \mathbf{a}_{t,i}) + \mathbf{B}_{-2}^{c,a}(\mathbf{c}_{t,i,j-2} \otimes \mathbf{a}_{t,i}) +$$

$$\mathbf{B}^{c,a}_{-1,-2}(\mathbf{c}_{t,i,j-1} \otimes \mathbf{c}_{t,i,j-2} \otimes \mathbf{a}_{t,i})]. \qquad (4)$$

where the matrices $\mathbf{b}_0^c$, $\mathbf{B}_{-2}^c$, $\mathbf{B}_{-1}^c$, $\mathbf{B}^a$, $\mathbf{B}_{-1,-2}^c$, $\mathbf{B}_{-1}^{c,a}$, $\mathbf{B}_{-2}^{c,a}$, $\mathbf{B}_{-1,-2}^{c,a}$ are constants which specify the functional form of V_a.

The *word potential function* $V_w : D_w \times D_c \to \mathcal{R}$ for the MRF is defined by the formula:

$$V_w(\mathbf{c}_{t,i,j}, \mathbf{w}_{t,i,j}) = \mathbf{c}_{t,i,j}^T[\mathbf{B}^{c,w}\mathbf{w}_{t,i,j}]. \qquad (5)$$

where the matrix $\mathbf{B}^{c,w}$ contains the constants which specify the functional form of V_a.

The *local word concept probability mass function* $p_c : D_c \to [0, 1]$ for the ASMURF is defined such that:

$$p_c(\mathbf{c}_{t,i,j} | \mathbf{c}_{t,i,j-2}, \mathbf{c}_{t,i,j-1}, \mathbf{c}_{t,i,j+1}, \mathbf{c}_{t,i,j+2}, \mathbf{a}_{t,i}, \mathbf{w}_{t,i,j}) =$$

$$\frac{exp(\phi_{\mathbf{c}_{t,i,j}})}{\sum_{\mathbf{c}\in D_c} exp(\phi_{\mathbf{c}})}$$

where

$$\phi_{\mathbf{c}_{t,i,j}} = V_c(\mathbf{c}_{t,i,j-2}, \mathbf{c}_{t,i,j-1}, \mathbf{c}_{t,i,j}, \mathbf{a}_{t,i}) +$$

$$V_c(\mathbf{c}_{t,i,j-1}, \mathbf{c}_{t,i,j}, \mathbf{c}_{t,i,j+1}, \mathbf{a}_{t,i}) +$$

$$V_c(\mathbf{c}_{t,i,j}, \mathbf{c}_{t,i,j+1}, \mathbf{c}_{t,i,j+2}, \mathbf{a}_{t,i}) + V_w(\mathbf{c}_{t,i,j}, \mathbf{w}_{t,i,j})$$

III. Theoretical Results

A. Representation Problem

Theorem 1: ASMURF Representation Theorem Consider a ASMURF $(\tilde{\mathbf{f}}, \tilde{\mathbf{a}}, \tilde{\mathbf{c}}, \tilde{\mathbf{w}})$ with neighborhood graph G such that:

- $\{\tilde{\mathbf{f}}_{t-2}, \tilde{\mathbf{f}}_{t-1}, \tilde{\mathbf{f}}_t\}$
- $\{\tilde{\mathbf{a}}_{i,1}, \ldots, \tilde{\mathbf{a}}_{i,d_i}, \tilde{\mathbf{f}}_i\}$, and
- $\{\tilde{\mathbf{c}}_{i,j,1}, \ldots, \tilde{\mathbf{c}}_{i,j,d_{i,j}}, \tilde{\mathbf{a}}_{i,j}\}$

are the only cliques of G for $t = 3, \ldots, d$, $i = 1, \ldots, d$, and $j = 1, \ldots, d_i$. The joint distribution of $(\tilde{\mathbf{f}}, \tilde{\mathbf{a}}, \tilde{\mathbf{c}}, \tilde{\mathbf{w}})$ may be represented by the probability mass function in (1) without any loss in generality.

Proof: Inspection of the particular parametric form of V_f, V_a, V_c, and V_w presented in Theorem 1 shows that any arbitrary choice of V_f, V_a, V_c, and V_w may be represented. The proof of the theorem then follows directly from the Hammersley-Clifford Theorem (see Besag, 1974; Geman & Geman, 1984; Golden, 1996; Winkler, 1991, for relevant reviews). ■

B. Learning Problem

Correlation matrices (e.g., $\mathbf{B}^f_{-2}$ or $\mathbf{B}^{c,a}_{-1}$) are estimated using the method of moments. For example, $\mathbf{B}^f_{-2}$ is estimated by:

$$\hat{\mathbf{B}}^f_{-2} = n^{-1} \sum_{s=1}^{n} \mathbf{f}^s_t [\mathbf{f}^s_{t-2}]^T$$

and $\mathbf{B}^{c,a}_{-1}$ is estimated by:

$$\hat{\mathbf{B}}^{c,a}_{-1} = n^{-1} \sum_{s=1}^{n} [\mathbf{c}^s_{t,i,j}](\mathbf{c}^s_{t,i,j-1} \otimes \mathbf{a}^s_{t,i})^T.$$

From a computational perspective, efficiency is greatly enhanced because the resulting correlation matrices will tend to have many zero elements so sparse matrix representation and manipulation methods may be fully exploited.

Define the *logical count function* $\mathcal{L}_\theta : \mathcal{R}^{m \times n} \to \mathcal{R}^{m \times n}$ be defined such that the ijth element of $\mathcal{L}_\theta(\mathbf{M})$ is equal to 1 if the ijth element of $\mathbf{M}$ exceeds some *minimum count threshold* θ and the ijth element of $\mathcal{L}_\theta(\mathbf{M})$ is equal to 0 otherwise. Let m_{ij} bs the ijth element of $\mathbf{M}$. Now define the *log-log count function* $\log\log_\theta : \mathcal{R}^{m \times n} \to \mathcal{R}^{m \times n}$ be defined such that the ijth element of $\log\log_\theta(\mathbf{M})$ is equal to $log(log(m_{ij}))$ when $m_{ij} > \theta$ and is equal to zero otherwise. The log-log count function is inspired by Zipf's law which states that count frequencies in the language domain tend to follow a log(log) distribution.

In practice, the following two learning rules have been found to be effective. We define them by example. Estimating $\mathbf{B}^{c,a}_{-1}$ using the formula:

$$\hat{\mathbf{B}}^{c,a}_{-1} = \sum_{s=1}^{n} \mathcal{L}_\theta \left([\mathbf{c}^s_{t,i,j}](\mathbf{c}^s_{t,i,j-1} \otimes \mathbf{a}^s_{t,i})^T \right)$$

will be referred to as a *logic* learning rule whose *minimum cell count* is θ. Estimating $\mathbf{B}^{c,a}_{-1}$ using the formula:

$$\hat{\mathbf{B}}^{c,a}_{-1} = \sum_{s=1}^{n} \log\log_\theta \left([\mathbf{c}^s_{t,i,j}](\mathbf{c}^s_{t,i,j-1} \otimes \mathbf{a}^s_{t,i})^T \right)$$

will be referred to as a *log log weighted* learning rule whose *minimum cell count* is θ. A useful choice for the minimum cell count threshold value is $\theta = 2$ to avoid over-fitting.

C. Inference Problem

In practice, the segmentation strategy is not known but a heuristic algorithm which examines all possible segmentation strategies for a single molecular proposition may be used. The heuristic algorithm applies all possible segmentation strategies to the first group of words and selects the segmentation strategy yielding the most probable molecular proposition associated with the first word group. Then, the most probable first molecular proposition and segmentation strategy generated by the system is assumed to be correct and the heuristic algorithm only needs to examine all possible segmentation strategies for computing the second molecular proposition. This incremental process then continues until the automatic semantic annotation process is complete. Besag's (1986) Iterated Conditional Modes (ICM) algorithm was used to compute a suboptimal solution to the problem of finding maximum a posteriori estimates of the semantic annotation label values. The inference problem is further simplified through the use of the following heuristic. In situations where a word and a word-concept never co-occur during the learning process, it is assumed that they can never co-occur. Similarly, in situations where an atomic proposition and word-concept never co-occur during the learning process, it is assumed that they can never co-occur. And finally, in situations where a complex proposition and atomic proposition never co-occur during the learning process, it is assumed that they can never co-occur.