Dynamic Context Generation for Natural Language Understanding: A Multifaceted Knowledge Approach

Samuel W. K. Chan, *Associate Member, IEEE,* and James Franklin

***Abstract*—We describe a comprehensive framework for text understanding, based on the representation of context. It is designed to serve as a representation of semantics for the full range of interpretive and inferential needs of general natural language processing. Its most distinctive feature is its uniform representation of the various simple and independent linguistic sources that play a role in determining meaning: lexical associations, syntactic restrictions, case-role expectations, and most importantly, contextual effects. Compositional syntactic structure from a shallow parsing is represented in a neural net-based associative memory, where it then interacts through a Bayesian network with semantic associations and the context or "gist" of the passage carried forward from preceding sentences. Experiments with more than 2000 sentences in different languages are included.**

***Index Terms*—Connectionism, context-dependent model, knowledge representation, natural language understanding.**

I. Introduction

CONTEXT effects have always been difficult for natural language understanding projects that follow the traditional plan of dealing with syntax before semantics. While many classical linguists have claimed that syntax should play the primary role [1], a number of now well-known phenomena show that humans use semantic associations and their understanding of context to assist with, make up for, or even override syntax [2]. Texts grossly ill-formed grammatically can be understood, apparently by much the same processes as are used for well-formed texts. An utterance like "*woman, street, crowd, traffic, noise, thief, bag, loss, scream, police*" is grammatically almost structureless, but the associations common to the words indicate a particular narrative. The same effect is relied on by some of the successes of natural language processing (NLP), such as the programs that translate web pages: the reader is expected to fill in the semantic gaps in the results, which are generally ill-formed both syntactically and semantically. The bible of the syntactic approach, Chomsky's *Syntactic Structures*, instances "*Read you a book on modern music?*" as a paradigm of an ungrammatical sentence [3], but the same speakers who recognize it as ill-formed have no trouble understanding it. Many of the classical examples that illustrate types of ambiguity also show that semantics is processed as early as syntax and can steer the syntactic processing. "*The chickens are ready to eat*" is ambiguous (whereas, neither "*The cakes are ready to eat*" nor "*The diners are ready to eat*" is ambiguous). This depends on the fact that chickens can both eat and be eaten; the two possibilities lead to two different decisions on the grammar of the sentence. An NLP system can only begin to address this problem by activating simultaneous chains of associations that can lead to the two possible connections between chickens and eating. The same applies to "*He saw the girl with a periscope*." The semantic association between periscopes and vision must be active to prevent the sentence being parsed in the same way as "*He saw the girl with a dog*." In addition, the need for semantic considerations in anaphora resolution is well-known and illustrated by such sentences as "*She dropped the glass on the floor and it broke*." While reference to specially encoded world knowledge about relative fragilities is always possible in such cases, an association between glass and breaking is more psychologically plausible. In view of the tendency of short words to have many meanings, it is clear that the simultaneous activation of associations is necessary for disambiguation when the true meaning relies on matching of common associations.

The lesson that semantic associations ought to be active at the same time as syntactic processing is reinforced when we come to consider the effects of context. There are two kinds of context effects, both of which indicate, in different ways, the need for a network of associations to process natural language. The first is the context of understanding of the total gist of a passage, which is carried forward into the interpretation of new sentences. In "*My sister's kids are at the store again. Those three boys sure like candy*," there must be an association (of near-synonymity) active between "*kids*" and "*boys*" to enable the identification of the two. Less widely appreciated is the importance of semantic associations in setting up a context of interpretation at the very beginning of a text. Writers of a text need to begin with something that will very quickly show the reader what sort of subject matter is being talked about. If a text begins: "*Waiter!*," "*Sir?*" the reader can infer the restaurant context immediately from the associations of the first word, and the relation between the speakers in the dialogue is revealed by the first two words. The reader knows to expect a next sentence like "*There's a fly in my soup*" (and not one like "*We ride at dawn*"). We therefore agree with those approaches which try to generate a model of context that embodies the meaning of both sentences and whole texts [4].

Although there is some understanding and local consensus on the impact of the semantic theory on context generation, there are barriers to transforming the semantic information into con-

Manuscript received April 2, 1999; revised November 5, 2001. This work was supported in part by two grants from the Research Grants Council of the Hong Kong Special Administrative Region under Projects CUHK 1196/99E and CUHK 1221/00E. This paper was recommended by Associate Editor L. O. Hall.

S. W. K. Chan is with the Department of Decision Sciences, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (e-mail: swkchan@cuhk.edu.hk).

J. Franklin is with the School of Mathematics, University of New South Wales, Sydney, NSW 2052, Australia.

Digital Object Identifier 10.1109/TSMCA.2003.811129

text representation [5]. The main hindrance is that, compared to syntactic analysis, semantic theory is less precise even under general underlying principles of semantics in a restricted domain [6]. Context generation for the understanding of sentences is usually split into two stages [7], [8]. First, it is necessary to determine the appropriate meaning of each word. Lexical items that are indexed by their dictionary definitions and thesaural categories are well-known to readers. Associations, such as the *bank/money* pair, not only provide interrelations between two pieces of explicitly stated information which may contribute to the construction of inferred connections, they also provide frames of references through which information never stated in a text can also be acquired [9]. In the second stage, with respect to the context in which the sentence is used, every reader tries to integrate every encountered lexical item with the forgoing text. If the lexical item is compatible with the context, the interpretation process will proceed. These prior context effects facilitate current language interpretation by suppressing the irrelevant possible interpretations; thus, it helps to narrow down the semantic inferences to the most plausible one.

This paper attempts to address some of the problems involved in dynamic context generation in natural language understanding. We describe a context learning model which supports a flexible and coherent interpretation of utterances. The model is based on a connectionist architecture for learning context-dependent representation. We illustrate how the multiple informative linguistic constraints and their interaction, which encompasses both symbolic and connectionist reasoning, can be modeled in order to assemble the dynamic context generation. Various information including syntactic, semantic, and context knowledge are taken account of simultaneously, rather than in separate stages, even though they are all different and orthogonal in nature, i.e., the information provided by each source is independent of and not directly relevant to any of the others. None of these knowledge sources will fully disambiguate the context or provide a self-sustainable solution to the problem of language understanding, but each will provide many clues. In this paper, we discuss the representation of each linguistic source, describe the dynamic context generation mechanism, and report on an experimental prototype. However, it is not the main objective of this paper to build all knowledge sources but to model the linguistic association which is fundamental to the learning process as well as being the essence of semantics. While we take advantage of the essentially continuous subsymbolic information about strengths of associations among the linguistic knowledge sources in our context generation, we attempt to reduce the discrete syntactic component but include more information in lexical entries in order to reduce the distinction between syntax and lexicon. All the linguistic information is represented in a uniform format so that the system can interact with the information from associations.

The paper is organized as follows. After the discussion of related work in Section II, we will discuss the formation of each knowledge source. In particular, we explain thoroughly how our system can make use of the linguistic clues spanning more than a single knowledge source. In Section III, we describe how lexical primitives can be generated in a recursive auto-association memory (RAAM). The lexical primitives fit together to capture underlying semantic relations and to provide the complement of syntactic analysis. The recursive compositional nature of the surface syntactic structure of sentences is formulated in a tree structure, as described in Section IV, with syntactic primitives as its nodes. The syntactic primitives provide a means for integrating with lexical primitives according to the syntax of the language, even though they are heterogeneous in nature. Section V shows how a semantic resolution process can distill the sentential meaning from the multifaceted knowledge sources. Another major characteristic of our system is to model the context effects in semantic interpretation. In Section VI, we shall explain how our context generation can be achieved and how the system integrates context clues in the preceding sentences. In order to demonstrate the capability of our system, several simulation experiments are delineated in Section VII. A full evaluation of our system is discussed in Section VIII followed by a conclusion.

II. Related Work

One of the most pervasive phenomena in natural language is that of semantic interpretation. This problem confronts language learners and natural language processing systems alike. As in work on syntactic processing, theoretical linguistic research on semantic interpretation is conceptualized almost entirely within symbolic frameworks which exploit the use of logic and frames in text understanding [10]–[12]. While predicate logic provides very precise inferences, without any tolerance, when all the linguistic preferences are enumerated in advance, it suffers significantly since linguistic information is most unlikely to be complete in many real-world language understanding problems. As demonstrated in a well-known treatise [13], the inherent difficulty of semantic interpretation of the word *pen* in both *The box is in the pen* and *The pen is in the box* seems impossible to solve solely in any theoretical logic-based language system. Indeed, theoretical symbolic-based linguists are fully aware of the daunting task of the scale-up problem in semantic interpretation.

One of the recent approaches to overcome the brittleness of symbolic natural language understanding processing aims at releasing various linguistic markers for limited concepts of interest in a connectionist paradigm. Cottrell and Small developed a connectionist architecture in word sense disambiguation [14]. Every single unit of their connectionist network is inter-connected at three linguistically significant levels: a lexical, a word sense, and a case logic level. A sentence is fed into the network at the lowest lexical level as provided by the lexicon and the grammatical morphemes of a language. Activation is then spread among the associated word sense and the nodes at the case logic level. Understanding a linguistic expression is the result of multiple computational cycles which settle into a stable pattern. Waltz and Pollack [15] and Lange [16] describe two similar approaches with an explicit syntactic representation structure of a sentence. St. John and McClelland present a connectionist model which learns to assign semantic representations to English-like sentences [17]. The task of the model is to process a single-clause sentence into a representation of the event it describes. Although the details of their model are somewhat complex, the main idea is that, via association, a network is trained to produce a correct semantic representation of the situation described by each input sentence.

All the connectionist approaches employ propagation mechanisms in their work even though their systems are fragile due to no prior knowledge being incorporated [18]. Miikkulainen has tried to remedy the situation by presenting a connectionist model which reads partial script-based stories and paraphrases them as causally complete output stories using distributed representations [19]. Episodic memory is used to store hierarchical script representations, in which the top level represents the script while traces and specific instantiations are encoded in the lower level. The model is trained only on pre-parsed script structure whose syntactic structure has already been analyzed. Whatever disambiguation is necessary for processing the stories is done only by the sentence parser. No explicit context generation and disambiguation mechanism has been actually implemented and there is no attempt to model contextual effect.

Another approach to overcoming the brittleness in symbolic natural language understanding is to acquire linguistic knowledge from large corpora. Statistical methods gained popularity because of the predictive model of language through the extensive analysis of word patterns and their ability to in infer semantic information from the observed distribution of words. Current statistical methods allow systems to acquire large quantities of high quality general-purpose knowledge, which practically eliminates costly and error-prone hard coding in pure syntactical linguistic processing. Various applications have proved to be sophisticated, such as in syntactic [20], semantic [21], part-of-speech [22] tasks, and in the acquisition of taxonomic knowledge [23]. In contrast to in-depth natural language understanding tasks, Cardie uses standard symbolic machine learning algorithms, i.e., decision tree induction and the k-nearest-neighbor algorithm, to identify the trigger word for an extraction pattern, the general linguistic context in which the patterns would be applied [24]. However, the attitude has been criticized in that no conceptual description is provided and it is generally restricted to a *shallow* understanding [25]. Distribution statistics are not sufficient by themselves in any semantic studies and they suffer from, at least, two major obstacles in deep natural language understanding. The first obstacle is that such statistics concern the distribution of words, whereas the semantic theory of distribution concerns the distribution of word-senses. The second obstacle to extracting semantic information from co-occurrence statistics is that nonsemantic factors, such as the use of pronouns, can influence the choice of words and thus the distribution of sense-uses.

A different procedure for incorporating relevant linguistic knowledge, without human intervention, aims at making use of some existing lexical databases, such as WordNet [26]. Researchers in natural language understanding have viewed lexical databases as a means of investigating the semantic structure of natural language as well as resources for overcoming the bottleneck in knowledge acquisition. Different approaches have tried to integrate diverse sets of knowledge sources to disambiguate word sense with WordNet [27], [28]. Without any text inference, they employ supervised learning from a set of tagged sentences. Harabagiu and Moldovan provide an attempt at text inference using WordNet [29]. They devise a parallel technique for building network-based information agents which are capable of retrieving texts having high contextual similarity. Their main assertion is that a pair of texts having the largest number of common concepts in the same semantic constructs tend to have closer contexts. Instead of providing the context and generating a context-dependent representation for paraphrasing the subsequent sentences and maintaining a hierarchical organization on context according to the underlying structure of utterances, they devise a coherence metric that captures the common concepts and lexical relation spanning any pair of texts.

In the research of context, the general approach in acquiring context aims at utilizing different classifiers which are generated from different shallow knowledge bases. Resnik investigates four different methods for combining three shallow knowledge sources of context patterns [30]. In order to resolve syntactic ambiguities, the context patterns are compared term by term, from most reliable to least reliable, until some match is found. Similarly, Yarowsky find local contexts to be a powerful sense indicator [21]. He suggests choosing a sense by matching a set of context patterns in word sense disambiguation. Two different types of context (*topical* and *local* context) have already been distinguished [31]. The topical context is comprised of substantive words that are likely to co-occur with a given sense of other words. Topical context is generally insensitive to the order of words. The local context includes information on word order and syntactic structure. They argue that topical context alone is not sufficient for text information retrieval. They show the addition of local context improves the overall performance. None of these attempts try to define precisely the concept of linguistic context. They account for context simply by using windows of surrounding words in tagging, such as in word-sense disambiguation.

In the direction of deep reasoning in text inference, McRoy and her colleagues propose a mixed-depth representation for dialog processing [32]. One of the characteristics of their approach is to allow a single representation framework to produce a detailed representation of requests. Their mixed-depth representation serves as a central blackboard which encodes both syntactic and conceptual information. Although the application of the blackboard in the learning of contextual effects, which is the main focus of their paper, is still under exploration, their approach may provide certain clues and techniques which will be of value in knowledge organization in linguistic reasoning. Certainly, sentential meanings are not solely determined by their word patterns but also rely on their lexical knowledge, case-role expectations, syntactic structures, semantic associations, collocation, or more importantly, context effects. While most long-range research efforts in artificial intelligence are trying to understand the meaning of sentences by utilizing these items of knowledge individually, all these knowledge sources should be acquired in parallel in order to capture sentential meaning in semantic interpretation [33]–[35].

As can be seen above, although the sophistication of natural language processing methods may vary from simple tagging to deep understanding, all these approaches demonstrate the importance of integrating and manipulating contextual knowledge into language understanding as well as the impact of different knowledge sources in context formation. The next logical step is to generate the context representation during comprehension using a diversity of knowledge sources and make use of them in the text understanding. Our system takes a different approach from most of the research efforts described above by generating a dynamic context representation. Our dynamic context generation is on the basis of sentence meaning at multiple levels. We will show how each separate, but complete structure for syntactic, semantic and, in particular, the contextual knowledge can be encoded and generated. The main philosophy is that natural language utterances are always interpreted in some context. Our model of natural language explicitly represents context and provides a mechanism for processing it. Our context modeling is handled by the same representational and inferential mechanism which allows us to impose context-dependent constraints while interpreting natural language sentences.

Although integrating multifaceted knowledge and local-context-based methods for processing large corpora is not a new idea, the basic tenet of our approach is that our context generation is not regarded as a separate issue from language understanding, but occurs simultaneously with the semantic interpretation. We argue that there is more to integrating the information from the different sources than juxtaposition. Another novel aspect of our approach is that we use a weak method to bootstrap the in-depth method. The weak method is based upon the hypothesis that two word-senses occurring in the same sentence will probably be semantically related. This prediction might be weak, but the existence of such connections is noncontroversial and essentially irrefutable. Most sentences in text are coherent wholes, and every pair of linguistic items is linked by some chain of relationships. In addition, with the aid of soft computing, semantic processing, or ambiguity, resolution is deferred until all the relevant knowledge is available. Every linguistic item is integrated with the subsequent text in our dynamic context. The context generated facilitates language understanding by suppressing the irrelevant concepts. The capability of the context representation to narrow down the semantic inferences provides the basis for the development of a comprehensive language learning model. In other words, our dynamic context generation is specially designed for knowledge acquisition from substantial portions of text, and that is what distinguishes it from other approaches. We are not aware of any approach that would resemble our limited-syntax, language-based knowledge representation to acquire, represent, and utilize linguistic knowledge from texts. The full evaluation of our system will further be illustrated in Section VIII.

III. Lexical Subsymbols as Primitives in Semantic Association

Lexical knowledge is of course one of the basic units of any language system [36]. Research in language understanding has in part concentrated on the development of suitable formalisms for expressing lexical information. While Elman has demonstrated that sequential context can provide the basis for adducing the category structure of internal representations of lexical items [37], we generate our lexical primitives mainly based on fields of meaning, using the Longman Lexicon of Contemporary English [38]. Each entry in the lexicon is in a bilingual format of both English and Chinese. The lexicon has 14 semantic fields of a pragmatic, everyday nature, with a simple index system for the ease of scaling up in meaning representation. Another unique feature of the lexicon is that many of the word sense definitions are marked with a subject field code (SF). The code signifies the subject area that the word-sense pertains to. For example, the *Money-Commerce*-related senses of *bank* are marked "J" and the senses can be further subdivided into sub-field (Sub-SF) *Banking* "E." Within each field and its sub-fields in the lexicon, the lexical items inter-relate and define each other in specific ways. On the other hand, there are many different types of linguistic items or parts-of-speech in language, such as verbs, nouns, adjectives, proper nouns, simple words, compound words, and even phrases, and such knowledge has been shown to be useful for various computational tasks such as language parsing and understanding. While the semantic category of an item and its part-of-speech reveal some of the surface and lexical relations and seem to be the main ingredients for constructing the lexical primitives, there is a strong tendency for language scientists to turn their focus onto some simple but robust indexes based on occurrences. Buoyed by the availability of corpora, we link a set of statistical attributes to our primitives in order to reflect their occurrences. Besides taking into account the relative frequency of lexical items as illustrated in the Academia Sinica Balanced Corpus 3.0 [39] which has more than 3 million tagged lexical items, an item saliency factor is another measure which computes the number of each item occurring in the corpus with respect to the number of texts in the corpus [40]. The saliency factor for each item $Sal(w_i)$ is defined as

$$Sal(w_i) = tf(w_i) \times \log \frac{N}{df(w_i)} \tag{1}$$

where item frequency $tf(w_i)$ is the number of occurrences of the item w_i in the corpus, while $df(w_i)$ is the number of texts in the corpus of N texts in which the items w_i occurs. Obviously, when N is large and $df(w_i)$ is small, the token w_i is considered to be more important than others. However, if N is large and the $df(w_i)$ is large too, the token w_i is considered to be less important in a corpus sense. The frequent items that are concentrated in particular texts are considered to be more important than equally frequent items that occur evenly over the corpus. In other words, items that commonly found throughout a collection are not necessarily good indicators of saliency. It is the more technical or context-specific words that best indicate the type of context. Let us illustrate the above discussion in the light of the following examples:

The notion of reduced representations was first introduced to allow a connectionist architecture to represent a compositional structure [41]. This is a fundamental problem in neural networks and has also been demonstrated in other related researches [42]–[45]. Effective examples of distributed reduced representations include the recursive auto-associative memory (RAAM) [42] and the holographic reduced representations [46]. In order to produce our lexical primitive for each lexical item as a reduced representation of the structural information

TABLE I
THREE DIFFERENT ATTRIBUTES OF LEXICAL ITEMS: EXTRACTED FROM THE ACADEMIA SINICA BALANCED CORPUS (3.0)

Lexical Item	Semantic Fields(SF)	Linguistic Objects (LO)	Corpus Attributes
Coffee	SF: Food (E) Sub-SF: General (A)	Noun; Simple Word	Relative Freq: 0.027 Saliency: 3.94
Current Account	SF: Money (J) Sub-SF: Investment (E)	Noun; Compound Word	Relative Freq: 0.064 Saliency: 19.6
Prefer	SF: Attitudes (F) Sub-SF: Liking (C)	Verb; Simple Word	Relative Freq: 0.038 Saliency: 2.18
Drowsy	SF: Body (B) Sub-SF: State (G)	Adjective; Simple Word	Relative Freq: 0.002 Saliency: 5.73

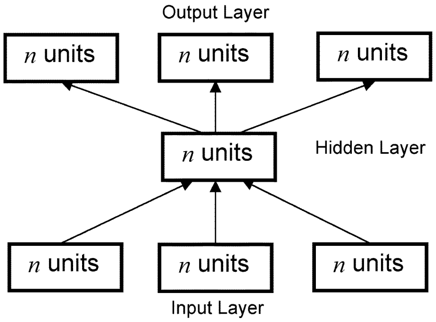


Fig. 1. General RAAM architecture which is defined for encoding and decoding trees of arbitrary branching factors.

as shown in Table I, a RAAM is used to encode the three different senses of the lexical items. The architecture of RAAM is essentially that of a simple three-layer network [47] which consists of equal-sized input and output layers, with a smaller hidden layer as shown in Fig. 1.

The network is trained by backpropagation to learn the identity function. It has been demonstrated that RAAM can be trained to reproduce patterns presented to its input layer on its output layer [48]. Merely reproducing patterns on an output layer is, of course, not very appealing. What is interesting is that such a network is able to encode the input pattern as an intermediate, condensed, distributed representation in the hidden layer. Using the auto-association technique, a RAAM can encode general tree structures of variable depth and fixed branching size into fixed-length distributed representations [42], [49]. The depth may be arbitrary in that no specific upper bound is placed on the depth of the trees encoded.

In the applications described in this paper, we use a slightly restricted version of the general RAAM, called a *tabular RAAM*. The tabular model is just like the general version, except that the data structures to be stored are in tabular format, rather than general trees. Since any sequence of elements in tabular format can be represented as a simple left-branching or right-branching binary tree, so the tabular model is really just a special case of the more general RAAM model. One advantage of the tabular model is that it allows us to employ a large number of processing units for the compressed hidden representations, in order to improve the information storage capacity without changing the representational scheme chosen for the tabular elements.

To gain some intuition into the types of representations formed by our tabular RAAM, we train a RAAM to encode the lexical primitives in the 25-20-25 RAAM, as shown in Fig. 2. It depicts the steps required to encode in the tabular RAAM.

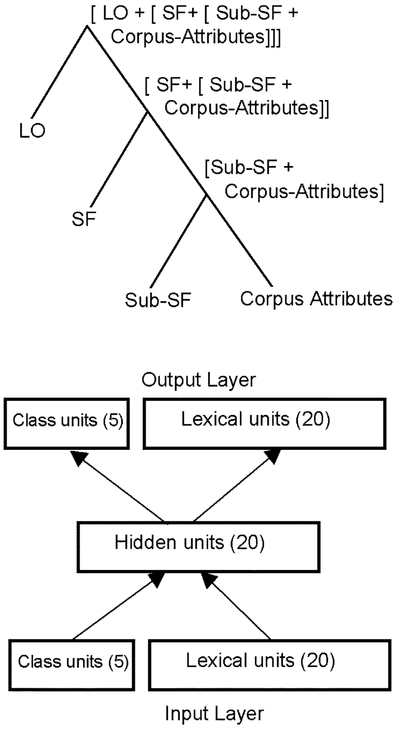


Fig. 2. Encoding lexical primitives in 25-20-25 RAAM.

To encode every lexical item, first, the corpus information, with eight bits for each attribute and the others are randomly set, is placed in the lexical units in the input layer while the subcategory of the subject field (Sub-SF) of the lexical items is assigned at the class units. Propagating these activations forward produces a compressed representation of [Sub-SF + Corpus-Attributes] on the hidden layer. Next, the representation of [Sub-SF + Corpus-Attributes] created in the previous step is placed in the lexical units with the its respective SF in the class units. Propagating these activations forward further produces a compressed representation of [SF + [Sub-SF + Corpus-Attributes]] on the hidden layer. Similarly, in the third encoding step, a compressed representation [LO + [SF + [Sub-SF + Corpus-Attributes]]] of the entire lexical item, as shown in Table I is obtained. Moreover, in order to minimize interference, each class unit is fixed with a minimum overlapping so as to keep the classes discriminated. An initial lexicon of 500 lexical items is used to train the RAAM with initial learning rate gradually changed from 0.1 to 0.01. The momentum rate of 0.75 is used. After the learning trials, the tabular RAAM's representational accuracy is tested by first encoding one of the items into a compressed representation and then immediately decoding that representation back into its constituent elements. If the decoded constituents match the original tabular elements, then the RAAM has learned to adequately represent the lexical items and its associated senses. As an initial test, 500 lexical items in the lexicon are recursively encoded and then decoded. Unsurprisingly, only four (out of 500 lexical items) cannot be decoded back to their corresponding input patterns and have an mean error greater than 0.05. The maximum error is 0.079.

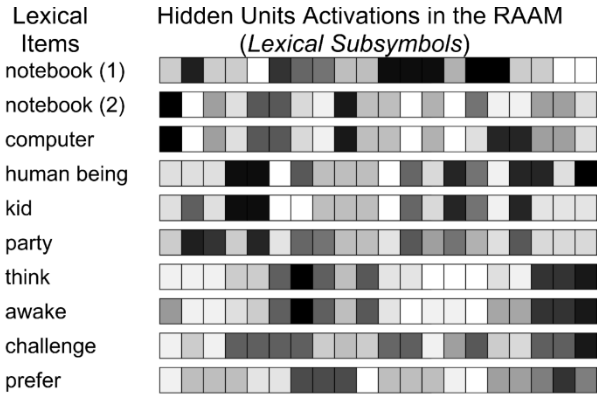


Fig. 3. Sample of lexical subsymbols so formed in the tabular RAAM.

The question to be asked is how this performance has been achieved. One way to answer this is to see what sorts of internal representation the memory develops. The internal representations are instantiated as activation patterns across the hidden units which are evoked in response to each relevant concept under the three basic senses. These patterns are saved at a training phase, during which no learning took place. Fig. 3 shows some of activation patterns in the hidden layer (hereafter, called *lexical subsymbols*) so formed in the 25-20-25 RAAM after 5500 epochs. Each of them is a 20-dimensional data vector which associates with its semantic meaning. It is most appropriate to regard the lexical subsymbols as sets of clues that constrain the meanings of the lexical items.

It can be observed that lexical items with similar meanings are represented using similar profiles, such as the lexical subsymbols of *notebook* and *computer* under `Machine`, but not the *notebook* under `Writing`. These lexical subsymbols are then subjected to hierarchical clustering analysis. Fig. 4 shows the tree constructed for more than hundred lexical items. The tree shows the similarity structure of the subsymbols and attempts to identify relatively homogeneous groups among them. Our RAAM has discovered that there are several major categories of lexical items. One large category corresponds to verbs in the lower half; another category corresponds to nouns. The noun category is divided into major groups for animates and inanimates which are then further subdivided. The lexical subsymbols show the essential *family-relationships* among the concepts. In most linguistic research, the notion of similarity plays a fundamental role in theories of linguistic knowledge and behavior. By means of our lexical subsymbols, it is possible to assess how similar two concepts are by observing the resemblance of the profiles between them. Each of the lexical subsymbols in these 20 dimensions forms a granule which is a fuzzy set of points having the form of a clump of lexical subsymbols drawn together by similarity. These lexical subsymbols allow a greater tolerance of errors in activation values. This is certainly not the case with conventional symbolic approaches to the representation of meanings. The semantic similarity measure between the lexical subsymbols is defined by

$$S(x_i, x_j) := \begin{cases} \langle x_i, x_j \rangle, & \text{if } \langle x_i, x_j \rangle \geq d_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

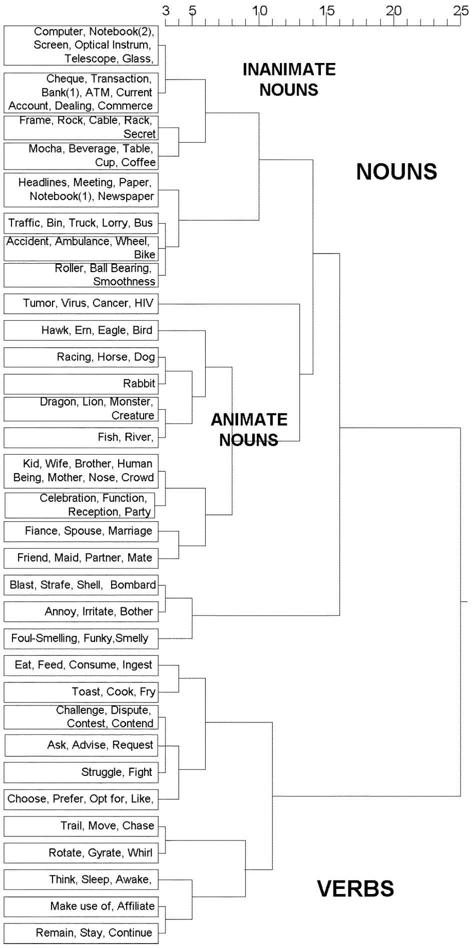

3
5
10
15
20
25
Computer, Notebook(2), Screen, Optical Instrum, Telescope, Glass,
Cheque, Transaction, Bank(1), ATM, Current Account, Dealing, Commerce
Frame, Rock, Cable, Rack, Secret
Mocha, Beverage, Table, Cup, Coffee
Headlines, Meeting, Paper, Notebook(1), Newspaper
Traffic, Bin, Truck, Lorry, Bus
Accident, Ambulance, Wheel, Bike
Roller, Ball Bearing, Smoothness
Tumor, Virus, Cancer, HIV
Hawk, Ern, Eagle, Bird
Racing, Horse, Dog
Rabbit
Dragon, Lion, Monster, Creature
Fish, River,
Kid, Wife, Brother, Human Being, Mother, Nose, Crowd
Celebration, Function, Reception, Party
Fiance, Spouse, Marriage
Friend, Maid, Partner, Mate
Blast, Strafe, Shell, Bombard
Annoy, Irritate, Bother
Foul-Smelling, Funky,Smelly
Eat, Feed, Consume, Ingest
Toast, Cook, Fry
Challenge, Dispute, Contest, Contend
Ask, Advise, Request
Struggle, Fight
Choose, Prefer, Opt for, Like,
Trail, Move, Chase
Rotate, Gyrate, Whirl
Think, Sleep, Awake,
Make use of, Affiliate
Remain, Stay, Continue
INANIMATE NOUNS
NOUNS
ANIMATE NOUNS
VERBS

Fig. 4. Hierarchical clustering analysis of lexical subsymbols which is formed from the activation patterns of the hidden units in the tabular RAAM.

where $\langle x_i, x_j \rangle$ is the dot product of the lexical subsymbols x_i, x_j and $d_{\max}$ is proportional to the number of lexical subsymbols in the system. In short, our lexical subsymbols serve as a means to reflect the lexical cohesion which contributes to the continuity of lexical meaning. The subsymbols provide the similarity not simply between pairs of lexical items but over a succession of a number of nearby related words spanning a topical unit of the utterance. Unlike other work in acquiring lexical meaning from machine-readable dictionaries [50], we generate our subsymbols based on their semantic classes and other corpus information. The lexical subsymbols, with their associations via similarity, in concert with other linguistic features, as discussed in the next sections, provide an easy-to-determine context to narrow down the search space to a specific meaning of a lexical item, in the resolution of ambiguities.

IV. Proposition Terms in the Syntactic Network

Although the lexical subsymbols provide some useful hints, other issues arise as we consider what sort of lexical representations might be required in sentences as shown in the following:

(S-1a) *Susan dropped the glass on the table.*
(S-1b) *Susan dropped the table on the glass.*

In a simple sense, the word “glass” may have the same meaning in the above two sentences. There may be controversy regarding

TABLE II
DIFFERENT TYPES OF p-TERMS

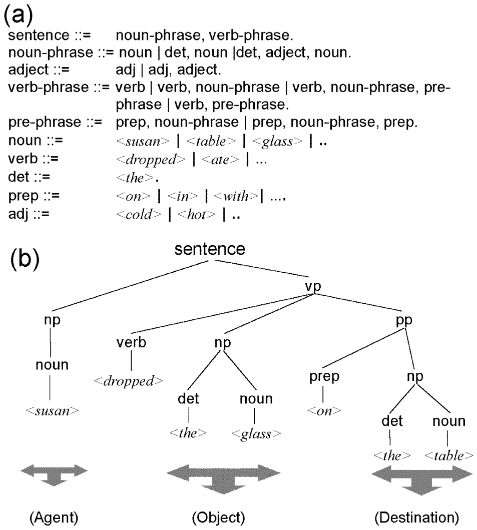
Type of p-terms	Example of p-terms	Denoted Concepts
Verb	DROP[AGENT[SUSAN], OBJECT[GLASS]]	*Susan dropped a glass*
Case-Role	OBJECT [DOCUMENT]	*Document as an object in Subject-Verb-Object (SVO) triple*
Adjective/Adverb	RIGHT [TIME]	*Right time*
Wh-	WHAT[RIGHT [TIME]]	*What is the right time?*
Modal	POSSIBLE [FATAL [FILE SAVING]]	*File saving may be fatal*
Atomic	ATOMIC [DOCUMENT]	*Document*

the appropriate way of representing the context-dependent role of the word "glass" in the above two sentences; however, it is clear the representations must somehow be different. The representations must reflect facts about "glass" in general as well as their usage in the specific contexts. These usage distinctions include the various roles filled by the "glass" and also the part of the sentence of which "glass" is a constituent. Fodor and Pylyshyn further argue that the ability to represent compositional relations in structured domains, such as language, is fundamental to a theory of cognition [51]. While it is notoriously difficult to represent the semantics of sentences, it is even more difficult, if not impossible, to represent the semantics without a system for binding arguments and their lexical meanings.

A syntactic structure of sentences indicates the way that lexical items in the sentences are related to each other. Unfortunately, a full syntactic analysis of every sentence in every text is too computationally demanding. In our approach, we employ a shallow parsing, based on case grammar, which does not require a full syntactic parse to pursue semantic analysis. The case grammar interpretation of a sentence is represented by a set of case frames whose slots are filled in by words in a sentence. This case frame representation is considered as a representation of the meaning of a sentence, so that it is comparatively independent of a particular language. That is, sentences in difference languages which express the same contents should normally have the same case frames [52]. This is the main reason why many machine translation systems have adopted the case frame representation as the final goal of the sentence analysis and the starting point of sentence generation. To represent the recursive structure of a sentence and the hypothesis as to its meaning after the parsing, the surface symbolic linguistic knowledge is organized using syntactic primitives, called *proposition-terms* (or simply *p-terms*). A *p-term* is a frame-like structure in the following format:

$$\mathtt{v}[\theta_1, \theta_2, \ldots, \theta_n]$$

where v is called the head concept label, and for each i, $1 \leq i \leq n$, θ_i is called a concept connector which may point to another p-term. Each constituent of a p-term is made up of a designated set of English token categories that commonly appear in the language, including `Verb`, `Case-Role`, `Adjective/Adverb`, `Wh-`, `Modal`, and `Atomic`. These p-terms offer a suitable basis for the inclusion of recursive syntactic information into language understanding. Table II shows an example under each type of p-term.


(a)

```
sentence ::=      noun-phrase, verb-phrase.
noun-phrase ::= noun | det, noun |det, adject, noun.
adject ::=        adj | adj, adject.
verb-phrase ::= verb | verb, noun-phrase | verb, noun-phrase, pre-
                  phrase | verb, pre-phrase.
pre-phrase ::=   prep, noun-phrase | prep, noun-phrase, prep.
noun ::=          <susan> | <table> | <glass> | ..
verb ::=          <dropped> | <ate> | ...
det ::=           <the>.
prep ::=          <on> | <in> | <with>| ....
adj ::=           <cold> | <hot> | ..
```

Fig. 5. (a) Sample of grammar rules used in the parser. (b) Parse tree with the syntactic constituents being fitted into the frame structure in order to generate the p-terms.

Given a set of grammar rules and a supporting lexicon, a syntactic parser is first employed in p-term identification in order to construct a parse tree [53]. The purpose of the syntax analysis is to organize and develop the sentence into syntactic units that play a role in determining its meaning. The parse tree provides the syntactic structure (or possible alternative syntactic structures) on which a semantic understanding system can rely. The meaning of the whole sentence is decomposed into p-terms starting from the main verb. Each verb will have a set of corresponding frames which is modified from the case structures proposed by Simmons [54].

The frames are stored in a frame repository. In other words, after the syntactic analysis is completed, our system generates the p-terms by fitting the syntactic entities into the frame structure of the main verb. We combine syntactic parsing and semantic pattern matching, using the sentence's syntactic structure and semantic frames, to generate the p-terms in order to represent the initial meaning of the sentence. Fig. 5 shows a sample of grammar rules used in the parser and the frame used in generating the p-terms. Obviously, some verbs may have more than one meaning, and the organization of p-terms may not be unique as Fig. 5 might indicate. In order to model the uncertainties in the organization of the p-terms, we employ a tree structure called a *syntactic network* (SN) in which the nodes of the network represent propositions or concepts. Each tree node represents a p-term in the set $\mathbf{X} = \{p_1, p_2, \ldots, p_n\}$ created in the identification phase. The syntactic network is motivated by its ability to represent probabilistic causal relationships among the p-terms and to constrain broad selections of grammatical features [55]. It is used to model the uncertainties involved in syntactic inferences as well as the grammatical encoding of the input sentences.

(S-2) *Rosalind walked to a store with her mum*

Fig. 6 shows how the syntactic structure of the sentence (S-2) can be represented in the syntactic network. The procedure for the incremental construction of the syntactic network is described explicitly in Algorithm A as follows.

```
[Algorithm A]
Objective: Incremental Construction of Syntactic Network
Goal: For every input sentence, set up a hierarchical structure which captures all the
  possible surface syntactic and semantic case structures of the sentence.
  • Subgoal [A.1] Identify the major fragments of syntactic structure of the input sentence
    [A.1.1] For each input sentence, parse the sentence into major syntactic constituents on
      the basis of the existing grammar rule set as shown in Fig. 5(a).
    [A.1.2] Generate a set of possible parsed propositions $PP$, from the parse tree by fitting
      the syntactic constituents into the frame structure by means of pattern matching
      as shown in Fig. 5(b).
  • Subgoal [A.2] Capture all the possible semantic case roles that are relevant to the
    input sentence.
    [A.2.1] For each element of the parsed propositions $p \in PP$, define a set of nodes,
      i.e., proposition nodes and concept nodes, $X_i \in \mathbf{X}$
    [A.2.2] Allocate the corresponding possible case roles of each concept element, as the
      child nodes, if any, in the parsed proposition $p$.
  • Subgoal [A.3] Setting up a syntactic network with uncertainty modeling
    [A.3.1] Append all the relevant concepts into $\mathbf{X}$ for each concept node and set up an
      ordering for all the concepts, $X_i \in \mathbf{X}$
    [A.3.2] Check whether there are concepts left in $\mathbf{X}$. In the affirmative case, do:
      [A.3.2.1] Pick a node $X_i$ and add a node to the network for it.
      [A.3.2.2] Set the parent node $\mathbf{P}(X_i)$ to some minimal set of nodes already in the net
        such that (3) is satisfied.
      [A.3.2.3] For all $i$, define the conditional probability table for $X_i$ which reflects
        the knowledge from the frame repository.
      [A.3.2.4] Zero conditional matrices enforce the mutual inhibitions. (This may arise
        when concepts, with both alternative meanings of a homonym, are constructed.)
      [A.3.2.5] Remove $X_i$ from $\mathbf{X}$.
```

Initially, for each input proposition, a set of all possibly relevant nodes is generated. Node p_1 in Fig. 6 stands for the main proposition. It is further divided into other p-terms $p_2, p_3, \ldots, p_6$. The syntactic decomposition is executed in the proposition parser which in turn parses the p-terms recursively. On the other hand, if the proposition is embedded into another proposition in a sentence

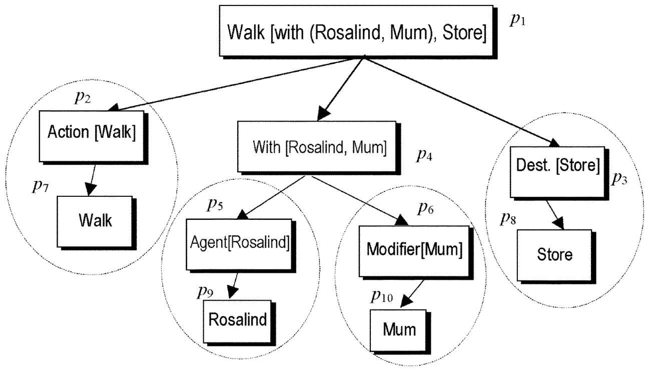


Fig. 6. Example of a syntactic network, illustrating the p-terms including proposition and concept nodes of the sentence (S-2).

(S-3) *Patrick thinks Rosalind walked to a store with her mum*

connections would be made to node p_1. Thus, these connections would indicate that the proposition node p_1 is the argument of the predicate *Think*. i.e., *Think*(*Patrick*, p_1). Fig. 6 should be viewed as a part of a much larger network. In addition, for each of the concept nodes, there is a case-role node associated with it. The organization of the nodes is not unique, particularly for the case-role nodes, as the figure might imply, but its spirit can be adopted for discussion purposes. Given this formalism, the conditional probabilities associated with input nodes indicate the possible roles and meanings for each concept and all probabilities can be calculated using the factorization of the joint probability distribution as shown in

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i \mid x_j\colon X_j \in \mathbf{P}(X_i)) \quad (3)$$

where x_i are possible values at node X_i, $P(x_1, \ldots, x_n)$ stands for $P(X_1 = x_1 \cdots$ and $X_n = x_n)$, and $\mathbf{P}(X_i)$ is the set of parent nodes of node X_i. From the standpoint of utilizing syntactic networks as a modeling tool in semantic interpretation, what needs to be done is to specify the parent nodes $\mathbf{P}(X_i)$, for each node X_i, and the conditional probability matrices associated with the links.

The linguistic information encoded in our syntactic networks facilitates the analysis of sentences by decomposing them into p-terms down to the concept level. The network captures the linguistic structures, and synthesizes sentence understanding under uncertainty. The syntactic network shows similarities to ACT in term of their move away from detailed analysis of syntactic structure and their degree of semantic abstraction [56]. At least two major characteristics are exhibited in this formalism. First, the syntactic network can in principle achieve invariance under paraphrase. This refers to the attempt to use a single representation for all sentences of identical meaning. The major advantage is that the conditional probability matrices need only be framed for one representation rather than for many different alternatives. Second, the uniqueness of p-terms in the syntactic network eliminates the exhaustive searches for the relevant information. Chunking of information in our syntactic networks

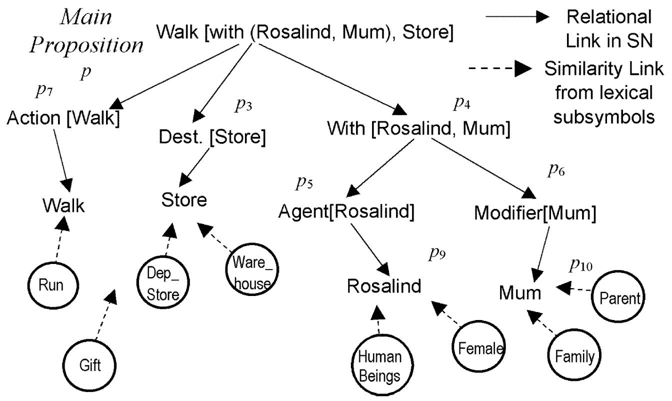


Fig. 7. Fragment of a syntactic network (SN). The circles indicate the lexical subsymbols activated by the semantic associations and are attracted by the relevant nodes in the SN based on the lexical similarity as defined in (2).

represents the knowledge about a p-term without being bogged down by the complexity of the concept. Although the formation of syntactic network is relatively complicated, the organization of p-terms into the syntactic network offers both the syntactic and semantic primitives in generating context representation as discussed in the next two sections.

V. Process of Semantic Resolution

Context-based language understanding relies on both the logic and coherence of a text, thus needing a representation of the context from which complex implications can be derived. Current work also shows that centrality of discourse should no longer be defined in terms of any simplistic linguistic rules, but rather in terms of linguistic ties which exist among discourse segments [57]. While our p-terms with the rigid formalism of syntactic analysis are organized in our syntactic network (SN), as discussed in previous section, in order to express composite and recursive properties of the linguistic constituents, syntactic structures are inadequate by themselves since they may lead to a blowout of computational complexity and also require knowledge of many probabilities that, in fact, are not known. Nor do they allow semantical considerations to bear on resolving syntactic ambiguities. Our lexical subsymbols, as discussed in Section III, provide the complement nicely where depth of reasoning is not required. Based on the similarity of the subsymbol profiles of the lexical items, for each incoming sentence, large assemblies of lexical subsymbols which are contiguous with the concepts mentioned in the p-terms of the syntactic network are spontaneously activated, as shown in Fig. 7.

In other words, while the activation of the main proposition in the syntactic network is dispersed via the relational links fastening the antecedent and consequent of the linguistic items, lexical subsymbols having high similarity measure will also be activated despite not being involved in the relational links in the syntactic network [58]. The lexico-semantic aspect of the lexical subsymbols account for the coherence implications derived by the text. In fact, these lexico-semantic effects in the subsymbols contribute the cohesion of the discourse. They have been identified as an important feature which underlines the structure of a unified text, distinguishing it from a nontext [59], [60]. In order to capture the syntactic as well as the lexico-semantic knowledge which may reflect a contingent of linguistic hypotheses about the input sentence, a knowledge matrix is derived as shown as follows.

```
[Algorithm B]
Objective: Generate a contingent of linguistic hypotheses about the input utterance
Goal: Identify the linguistic candidates and their inter-connections
  • Subgoal [B.1] Search along the Syntactic Network
    [B.1.1] Propagate the initial evidence vector, which is a unit vector, from the main $p$-term $p$ of the syntactic network down to its subordinates.
    [B.1.2] Initialize a corresponding initial evidence vector $(e_1, \ldots, e_n)$, for each extracted linguistic concept $p_i$, with the degree of belief $DB_i > t_c$, of the main $p$-term, where $t_c$ is an adjustable threshold. Each evidence vector serves as independent evidence, i.e.,
      [B.1.2.1] if $p_i$ is explicitly stated in $p$, $e_i \leftarrow 1$, else $e_j = 0$ when $i \neq j$.
      [B.1.2.2] if $p_i$ is an element in the syntactic network, $e_i \leftarrow DB_i$, else $e_j = 0$ when $i \neq j$.
  • Subgoal [B.2] Search along the Lexical Subsymbols
    [B.2.1] Activate some of the highly associated lexical subsymbols, for each extracted subordinate $p$-term $p_i$,
    [B.2.2] Define the evidence vector for all the activated lexical subsymbols $G_k$ i.e.,
      [B.2.2.1] if $G_k$ is the lexical subsymbol triggered by $p_i$, then $e_k \leftarrow \max(DB_i \times S(G_i, G_k))$ over all $i$, where $S(x, y)$ is a similarity measure defined in (2)
      [B.2.2.2] others are set to zero
  • Subgoal [B.3] Construction of the Knowledge Matrix $\mathbf{K}$
    [B.3.1] Propagate the evidence through the syntactic network, for each evidence vector.
    [B.3.2] Assign the resulting vector to the corresponding column of the knowledge matrix $\mathbf{K}$, if the maximal implication chains have been traversed.
    [B.3.3] Repeat until all the columns of the knowledge matrix $\mathbf{K}$ are complete.
```

The preceding algorithm has demonstrated how initial, enriched, but incoherent and possibly contradictory information collates into a single structure. It involves instantiating a set of elements corresponding to the input proposition by, first, setting evidence for the input proposition to unity and activating the related concept nodes from the syntactic network. Second, for each of these activated nodes with a high degree of belief, selecting a small number of its most closely associated neighbors from the lexical subsymbols with high similarity measures. Third, for the concept node i, calculate its impact on the concept node j through the SN. Fourth, for all pairs of nodes that have been generated, assign the calculated impacts into the knowledge matrix. The dimensionality of the knowledge matrix $\mathbf{K}$ is the number of all possible activated concepts, both propositions and subsymbols, of the input sentence. While the propagation of the evidence through the syntactic network is comparable to the style of the probabilistic reasoning as in the Bayesian network, each column of the knowledge matrix $\mathbf{K}$ represents the vector of initial relevance to all other activated concepts [55].

The formation of the knowledge matrix is local, associative, and without the guidance and control of any central agent. Activation is propagated in a massively parallel fashion from the main proposition down to the relevant lexical subsymbols. Obviously, under this mechanism, information which is less relevant, even sometimes contradictory to the input proposition, will also be extracted. As a result, not only a list of linguistic items is formed, but also, a fully interconnected network having positive and near-zero links is gradually created. Positive links between nodes may result, for instance, when two linguistic items are highly activated simultaneously. The near-zero ones may arise from a sentence where items are the two alternative meanings of a homonym. At the same time, this parallelism provides some linkages between two pieces of implicitly or explicitly stated linguistic constituents and is regarded as the process of linking phrase hypotheses generated from the SN into lexical hypotheses on the basis of lexical subsymbols. The likelihood of a lexical item to relate to another relies on the strength of links in the SN and the similarity of the lexical subsymbols. All these likelihoods are reflected in the knowledge matrix $\mathbf{K}$ which exhibits the lexical, syntactic and semantic clues about the sentence as its nodes. Moreover, the connections among them are interpreted as the degrees of *relevancy*. This structure accounts well for the similar parallelism and spontaneity in human reasoning processes [61]–[63]. The matrix $\mathbf{K}$ reflects blurred, ill-defined, and potential inferences, in the hope that some of them might turn out to be true. It comprises rough, piecemeal, and approximate facets of the input sentence which are then moulded into coherent wholes.

Our semantic resolution process takes the clues from disparate sources encoded in the matrix $\mathbf{K}$ which may contribute to context, and through cycles of competition, allows the best interpretation of a sentence to appear gradually as shown in Algorithm C below. Associated with each linguistic concept, either a node from the syntactic network or a lexical subsymbol, is a real number called its activation level which represents the strength of the concept during the semantic resolution. In other words, at some discrete time t, let vector $\mathbf{U}$ be defined by $\mathbf{U}(t) = (u_1, \ldots, u_n)$, where u_i is the activation level for the element i, and n is the number of elements extracted from the SN and the related lexical subsymbols. While the knowledge matrix $\mathbf{K}$ shows the likelihood that the linguistic concepts are activated by the input sentence and the degree of relevancy among the nodes, the activation vector $\mathbf{U}(t)$ represents the degree of *persistence* of the concept during the resolution. At the start, we assign to unity the nodes in $\mathbf{U}(0)$ with affirmative existence, such as the main proposition and the lexical items that are explicitly mentioned in the sentence, which is then passed into the knowledge matrix $\mathbf{K}$. Semantic resolution is interpreted as the repeated application of the function φ until convergence. For each $\mathbf{U}$, the vector $\varphi(\mathbf{U})$ is calculated which represents the updated activity. It is in effect an implementation of a constraint satisfaction process which spreads activation through the knowledge network $\mathbf{K}$. Each linguistic item in the knowledge network has some sort of activation value, i.e., central, important concepts are more tenaciously activated than peripheral ones. Continued spreading by repeated vector multiplication leads to equilibration. It strengthens the contextually appropriate elements and inhibits unrelated and inappropriate ones, so that smart and complex deductions can be achieved. The linguistic elements, which hang together in the network, strengthen or inhibit each other until a stable state is reached. The process stops at iteration m if $|\mathbf{U}(m) - \mathbf{U}(m-1)| < t_u$.

```
[Algorithm C]
Objective: Semantic resolution among the linguistic hypotheses generated from multi-faceted knowledge sources
Goal: Resolution by spreading activation in knowledge matrix $\mathbf{K}$
  • Subgoal [C.1] Network Initialization
    Set up the knowledge matrix $\mathbf{K}$ as described in Algorithm B.
  • Subgoal [C.2] Determine an initial activation vector for the semantic resolution
    [C.2.1] For the activation vector $\mathbf{U}(0) = (u_1, \ldots, u_n)$, assign unity activation values to the nodes with affirmative existence, such as the main proposition and the lexical items that are explicitly mentioned in the sentence, i.e., $u_i \leftarrow 1$
    [C.2.2] Set the others to zero in the vector $\mathbf{U}(0)$, i.e., $u_j \leftarrow 0$
  • Subgoal [C.3] Vector-matrix multiplication as a means of semantic resolution
    [C.3.1] Production firings direct the flow of activation from one element, the source, to another element through the knowledge matrix $\mathbf{K}$ and then are subjected to a normalizing operation. Mathematically, they are defined by the function, $\varphi$: $R^n \rightarrow R^n$, $\varphi(\mathbf{U}) = \varphi_2(\varphi_1(\mathbf{U}))$, where
      [C.3.1.1] $\varphi_1(\mathbf{U}) = \mathbf{UK}$, that is, $\varphi_1$ is a linear mapping given by multiplication by the matrix $\mathbf{K}$.
      [C.3.1.2] $\varphi_2(\mathbf{U}) = U/\sum_{i \leq n} |u_i|$
    [C.3.2] Continued excitation by repeated application of $\varphi$ leads to equilibration.
    The process stops at iteration $m$ if the $|\mathbf{U}(m) - \mathbf{U}(m-1)| \leq t_u$ where $t_u$ is the tolerance, another pre-set threshold which is used to control the accuracy of convergence in the process.
```

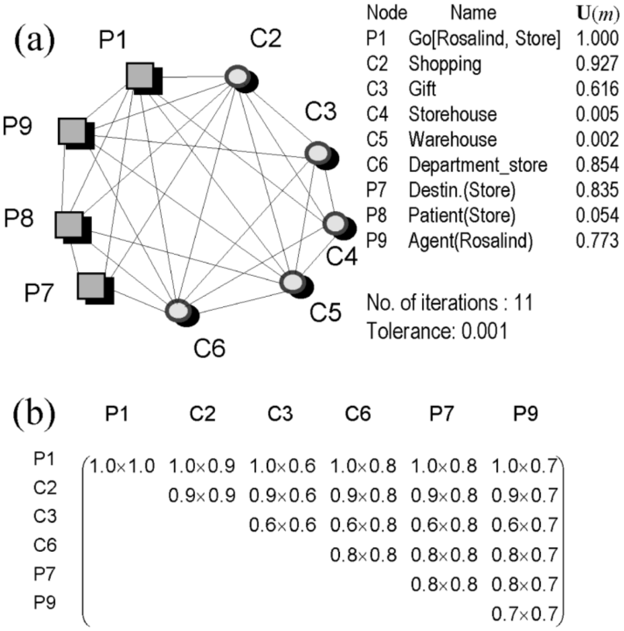
The final activation vector shows how strongly related items have strengthened each other, while unrelated or contradictory items have near zero activation values. In other words, the process is to reduce the dimensionality of the stimulus so that a very complicated stimulus could act as if only a small number of independent elements are involved. This process can be used to exclude unwanted elements from the matrix. They will have become deactivated in the semantic resolution process either because they are not strongly connected with the main part of the syntactic network or they are inhibited by activated nodes in the network. It is not difficult to observe that the system contemplates a large number of linguistic sources and then suppresses those which are irrelevant by virtue of its parallelism. It is completely different from the generate-and-test basis in any serial models used in the current paradigm of semantic interpretation.

VI. Generation of Context Representation

Throughout the literature of semantic interpretation runs a common theme: Language is understood in context. Linguists appeal to immediate linguistic context to give an account of the interpretation of ambiguous expressions. The only inferences coded automatically during interpretation are those based on easily available information and those required to make sentences in a text locally coherent. To acquire a correct interpretation of an utterance, the reader must first identify the general interpretation of each input sentence, as shown in previous sections, and combine these with the contextual assumptions generated in the preceding sentences to obtain the contextual effects. These contextual effects will govern the possible contextual implications for the utterance. As can be seen in the forgoing sections, for each input sentence, our system generates a vector representation, after the semantic resolution process, which contains all the distilled linguistic items. They are contextually relevant and all conflicts, as well as irrelevancies, have already been eliminated. To accomplish our objective of modeling context effects in understanding, we have adopted a proposition matrix Γ in which linguistic items are encoded and stored as patterns of interconnections between the items. More specifically, the linguistic items form a single, common proposition space which stores information about individual linguistic items. Rows and columns of the proposition matrix correspond to the distilled linguistic elements and the connections between them are represented by the nonzero entries. The matrix also specifies how strongly each distilled linguistic item in the sentence is related to every other. Unlike the knowledge matrix $\mathbf{K}$, the proposition matrix Γ is constructed from the asymptotic activation of each element. It represents the interrelations between elements remaining after the semantic resolution process. The strength of connection between elements is defined as

$$\mathbf{\Gamma}(r,\, s) = u_r \times u_s \tag{4}$$

where u_r, u_s are the final asymptotic activation values of the rth and sth linguistic elements after semantic resolution respectively. A given proposition matrix is a complex collection of information that contains the *distilled* elements, lexical, syntactic and semantic information, with each of them contributing a slice of the context of the sentence. The diagonal values of Γ represent the strength of linguistic elements in the space and the off-diagonal elements represent the strength of the relations between any two elements. The element with largest strength is said to be the *dominant node* of the sentence, while the others are called the *context nodes*.



Node	Name	$\mathbf{U}(m)$
P1	Go[Rosalind, Store]	1.000
C2	Shopping	0.927
C3	Gift	0.616
C4	Storehouse	0.005
C5	Warehouse	0.002
C6	Department_store	0.854
P7	Destin.(Store)	0.835
P8	Patient(Store)	0.054
P9	Agent(Rosalind)	0.773

No. of iterations : 11
Tolerance: 0.001

$$\begin{array}{c|cccccc} & \text{P1} & \text{C2} & \text{C3} & \text{C6} & \text{P7} & \text{P9} \\ \hline \text{P1} & 1.0\times1.0 & 1.0\times0.9 & 1.0\times0.6 & 1.0\times0.8 & 1.0\times0.8 & 1.0\times0.7 \\ \text{C2} & & 0.9\times0.9 & 0.9\times0.6 & 0.9\times0.8 & 0.9\times0.8 & 0.9\times0.7 \\ \text{C3} & & & 0.6\times0.6 & 0.6\times0.8 & 0.6\times0.8 & 0.6\times0.7 \\ \text{C6} & & & & 0.8\times0.8 & 0.8\times0.8 & 0.8\times0.7 \\ \text{P7} & & & & & 0.8\times0.8 & 0.8\times0.7 \\ \text{P9} & & & & & & 0.7\times0.7 \end{array}$$

Fig. 8. (a) After reading a sentence *Rosalind went to a store to buy a present*, a fully connected knowledge matrix $\mathbf{K}$ with rough, or even outright contradictory nodes is formed and is subjected to the semantic resolution process. The right column shows the asymptotic activation after the process. It is apparent that the incorrect meaning of *Store*, *Storehouse*, is deactivated. (b) After analyzing the sentence in the semantic resolution process, a proposition matrix Γ is formed containing the distilled linguistic elements.

The proposition *GO*[*Rosalind*, *Store*] in Fig. 8 is the dominant linguistic node with the largest strength $u_r \times u_r$ for the sentence. The figure also illustrates the connection strengths between the dominant node and its context nodes. For each input sentence, the proposition matrix Γ summarizes the inter-relations among the distilled elements constructed in the sentence [64], [65]. Based on the proposition matrices which reflect the interrelationships of the linguistic items from several sentences, the context effect on sentence j can be modeled in a context matrix M_j by the sum of proposition matrices, such as

$$M_j = \alpha_1\Gamma_{j-1} + \alpha_2\Gamma_{j-2} + \cdots + \alpha_k\Gamma_{j-k} \tag{5}$$

where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k$ to enforce a recency effect of the context.

The context effect is defined in the context matrix M_j which is constructed by summing the corresponding element from a set of k most recent proposition matrices in order to capture the recency effect in discourse understanding. Since the context matrix M_j and the knowledge matrix $\mathbf{K}$ are of the same structure, this feature gives an analytical simplicity in incorporating context cues into the subsequent sentence understanding and associative recall can be represented as the element-by-element matching. We formulate our discussion in applying the context matrix in modeling the context effect as follows.

Definition 1: $M^* = [m_{rs}]$ is the normalized matrix of $M = [m_{rs'}]$ where

$$m_{rs} = \begin{cases} 0, & \text{if } m_{rs'} \leq \beta \\ 1, & \text{otherwise.} \end{cases}$$

Definition 2: Two linguistic items P_r, P_s in a normalized context matrix $M^* = [m_{rs}]$ have path length n if P_r is connected to P_s in n moves, i.e., there are n nonzero paths between P_r and P_s in M^*.

Lemma 1: Let $[m_{rs}]^{(n)}$ be the (r, s)th element of M^{*n}. Then $[m_{rs}]^{(n)}$ is equal to the number of paths of length n from P_r to P_s.

Proof: Letting $m_{rs}^{(2)}$ be the (r, s)th element of M^{*2}, we have

$$(m_{rs})^2 = m_{r1}m_{1s} + m_{r2}m_{2s} + m_{r3}m_{3s} + \cdots + m_{rn}m_{ns}.$$

Now, if $m_{r1} = m_{1s} = 1$, there is a two-step connection, denoted as $P_r \rightarrow P_1 \rightarrow P_s$, from P_r to P_s. In other words, $P_r \rightarrow P_1 \rightarrow P_s$ is a two-step connection iff $m_{r1}m_{1s} = 1$. Similarly, for any value of $k = 1, 2, \ldots, n$, $P_r \rightarrow P_k \rightarrow P_s$ is a two-step connection from P_r to P_s iff the term $m_{rk}m_{ks} = 1$; otherwise, the term is zero. Therefore, m_{rs}^2 represents the number of 2-step connections from $P_r \rightarrow P_s$. A similar argument will work for finding the number of n-step connections from $P_r \rightarrow P_s$, thus $[m_{rs}]^{(n)}$ is equal to the number of paths of length n from P_r to P_s. ■

Definition 3: A subset of linguistic items in M_j^* is called a *context cue* Q_j if it satisfies the following three conditions.

i) The subset contains at least three linguistic items.
ii) For each pair of linguistic items P_r and P_s in the subset, both $P_r \rightarrow P_s$ and $P_s \rightarrow P_r$ are true.
iii) The subset is as large as possible; that is, it is not possible to add another linguistic item to the subset and still satisfy condition ii).

Obviously, Definition 3 means that the context cues Q_j are full maximal subsets which are self cohesive within the context matrix, even though M_j may be sparse. This cohesive cue provides a group of linguistic concepts (not only syntactic ones, which may concern semantic relations in the preceding sentences and which interrelates the substantive linguistic items). This cohesive cue may be aroused by either semantic signals or by vocabulary, i.e., by specific linguistic features and lexical expressions in the forthcoming sentences.

Definition 4: A context cue Q is retrieved by the sentence S if for any linguistic items $\alpha_Q \in Q$ and $\alpha_S \in S$, such that

$$\text{Max}\left(\frac{\sum \sigma^2(\alpha_Q, \alpha_S)}{\sum \sigma(\alpha_Q, \alpha_S)}\right)$$

where σ denotes a *similarity* relation over all linguistic items in a domain.

Lemma 2: Let $[m_{rs}]^{(3)}$ be the (r, s)th element of M^3. Then a linguistic element r belongs to some context clue Q if $[m_{rr}]^{(3)} \neq 0$.

Proof: If $[m_{rr}]^{(3)} \neq 0$, then there is at least one path of length 3 from linguistic item P_r to itself. Since $P_r \rightarrow P_r$ is not allowed, that is $[m_{rr}] = 0$, there must exist P_s and P_u such that $P_r \rightarrow P_s \rightarrow P_u \rightarrow P_r$. On the other hand, in our proposition matrix Γ, all directed links are two-way, so that we also have the connections $P_r \leftrightarrow P_s \leftrightarrow P_u \leftrightarrow P_r$. This means $\{P_r, P_s, P_u\}$ is either a context cue or a subset of a cue, as stated in Definition 3. ■

The preference calculation, as shown in Definition 4, identifies the potential context cues that may be relevant to the current analyzing sentence by the similarity measure σ. It calculates the semantic distance between the words in the input sentence and the context cue. Lemma 2 suggests a systematic procedure for retrieving the context cues that are relevant to the sentence in focus. By inspecting the nonzero diagonal entries of M^{*3}, the relevant context cues are invoked in the currently analyzed sentence. The relevant context cues are incorporated in semantic interpretation by superimposing Q_j into the current knowledge matrix K_j to form a new knowledge matrix K_j^* as follows.

$$K_j^* = K_j \oplus Q_j \tag{6}$$

where

$$\mathbf{Z} = \mathbf{X} \oplus \mathbf{Y} \text{ iff } z_{ij} = \begin{cases} x_{ij} + y_{ij}, & \text{if } i, j \text{ are concepts common to } \mathbf{X} \text{ and } \mathbf{Y} \\ x_{ij}, & \text{otherwise.} \end{cases}$$

Each activated context cue encodes some of the context knowledge that can be superimposed into the current knowledge matrix and carried over into the analysis as shown in (6). One of the most important implications in applying the context cues is that all the previously analyzed sentences behave like a group of experts in the current interpretation. Each of the proposition matrices inherits the factual and circumstantial facts from the corresponding sentence. Obviously, the resultant knowledge matrix is more robust than the individual knowledge matrices because the linguistic information is derived from a multiplicity of sources from the analyzed sentences, making the links in the current knowledge matrix less prone to error. In general, context effects may be drawn from a wide variety of information sources which may include the preceding text, cultural or scientific knowledge, or any item of shared or idiosyncratic information that the reader has access to at the time. It is worthwhile to mention that, in this architecture, we limit our scope of context effects by simply considering the preceding linguistic text but certainly not the environment brought to bear nor the set of nonlinguistic assumptions in which the utterance takes place. These are activated via the similarity relations of the subsymbols, as described earlier.

VII. Simulation Experiments

The primary testbed of semantic interpretation is the study of ambiguity. Any linguistic frameworks should justify themselves when an input sentence can be explained with more than one interpretation. In order to demonstrate the foregoing algorithms in more detail, in this section, we present a number of disambiguation phenomena which exhibit the ability of the framework to account for problems due to prepositional phrases and anaphora to supplement, verify, and strengthen our theoretical considerations above. With these major but different ambiguity issues and detailed discussions presented below, we will exhibit how our architecture advances the state of the art and stands out for its versatility as a semantic interpreter.

Modifiers such as adverbs and prepositional phrases cause a lot of problems in language understanding, simply because they can attach to several different heads. These are cases where multiple syntactic analyses can be assigned to a particular string of linguistic items. A typical structural ambiguity is shown in (S-4).

(S-4) *John saw the girl with the telescope*

Two possible interpretations are obtained simply because the prepositional phrase *with the telescope* can modify either the noun *John* or *the girl*. Obviously, there will have two distinct interpretations if the sentence falls into following different scenarios.

Scenario a): *John was looking out the window, trying to see how much he could see with various optical instruments. John saw the girl with the telescope.*

Scenario b): *Two girls were trying to identify some birds. One girl was holding a telescope and the other was not. John saw the girl with the telescope.*

This structural ambiguity can be resolved, in our architecture, by using context cues from the preceding sentences in order to arrive at a conceptually consistent interpretation. The two different interpretations of sentence (S-4) in the form of propositions can be shown as follows:

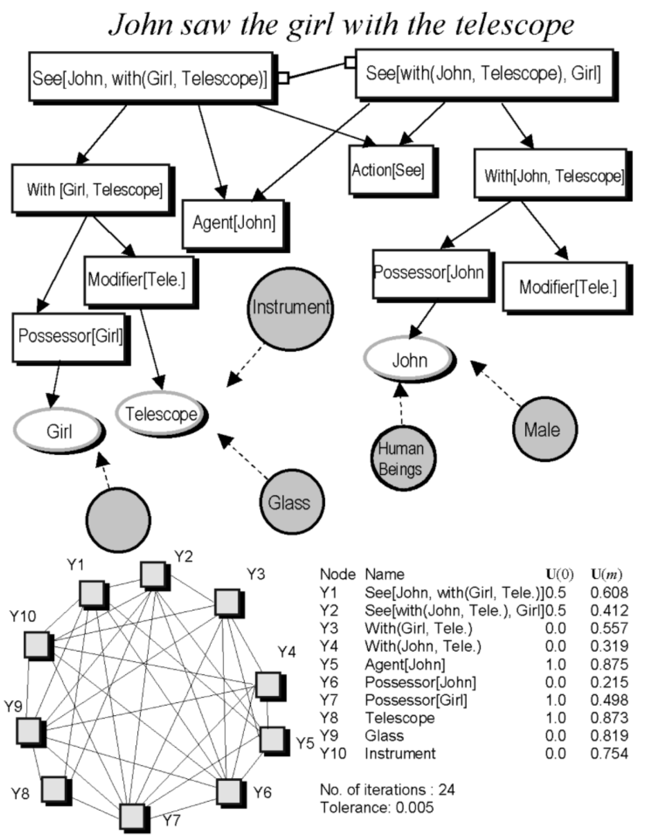
(S-4)* *See* [`Agent`: *John*, `Object`: *with* (*Girl*, `Instrument`: *Telescope*)]

(S-4)** *See*[`Agent`: *with* (*John*, `Instrument`: *Telescope*), `Object`: *Girl*]

The problem remaining is how the framework can distinguish the sentence (S-4) in the two different interpretations in a particular scenario, say, in scenario a). The sentence is first parsed in our parser and the possible p-terms are generated in semantic pattern matching. The organization of the syntactic network is on the basis of the conditional probability table which reflects the notion of causality as shown in Algorithm A. Fig. 9 shows the syntactic network of the sentence (S-4) with both of these interpretations activated in a parallel fashion. The two potential interpretations are connected by inhibitor links to ensure their noncoexistence in the final interpretation.

As can be seen in $\mathbf{U}(m)$ in the rightmost column of Fig. 9, the syntactic network and the subsymbols capture both the surface syntactic information and the underlying semantic meanings of the sentence. Although *See*[*John*, *with* (*Girl*, *Telescope*)] seems dominant with a higher activation value as shown in the figure, it cannot produce any decisive resolution in the interpretation. It is imperative for the forgoing context to be involved. The context matrix $\mathbf{M}$ that summarizes the interrelations among the distilled elements for scenario a) is shown in Fig. 10. Indeed, they contain the distilled information which is carried over into the current processing cycle.

In the simulation, due to the similarity between *optical instrument* and *glass* through the lexical subsymbols as shown in Fig. 4, the activation of *telescope* from the current sentence (S-4) brings in a full maximal context cue $Q \subset \mathbf{M}$. As defined in Lemma 2, the context cue Q is self-cohesive and represents the distilled information from the foregoing sentences after the preceding semantic resolution processes. In other words, the lexical item *telescope* brings influence to bear on the relevant context cue Q, urging the context cue to be involved into the semantic interpretation of (S-4) in scenario a). As a result, the linguistic items, such as *Agent*[*John*], *Instrument*[*Optical instrument*], *See* are superimposed into the current knowledge matrix $\mathbf{K}$ of sen-

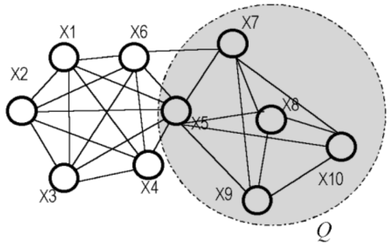


Node	Name	$\mathbf{U}(0)$	$\mathbf{U}(m)$
Y1	See[John, with(Girl, Tele.)]	0.5	0.608
Y2	See[with(John, Tele.), Girl]	0.5	0.412
Y3	With(Girl, Tele.)	0.0	0.557
Y4	With(John, Tele.)	0.0	0.319
Y5	Agent[John]	1.0	0.875
Y6	Possessor[John]	0.0	0.215
Y7	Possessor[Girl]	1.0	0.498
Y8	Telescope	1.0	0.873
Y9	Glass	0.0	0.819
Y10	Instrument	0.0	0.754

No. of iterations : 24
Tolerance: 0.005

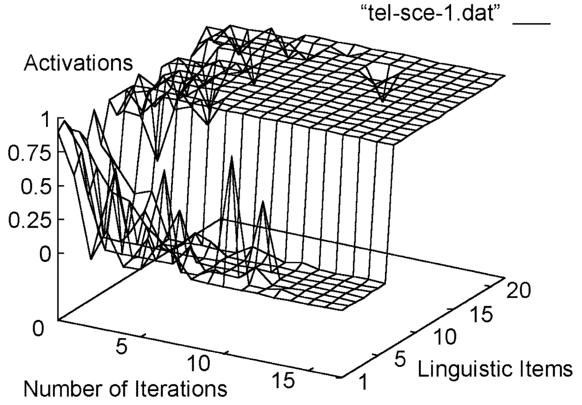
Fig. 9. Syntactic network for two possible interpretations of sentence (S-4) and a fragment of the knowledge matrix $\mathbf{K}$ so formed. The shaded circles are the subsymbols attracted by the corresponding concept nodes.

$$\mathbf{M} = \begin{array}{c} \begin{array}{cccccccccc} X1 & X2 & X3 & X4 & X5 & X6 & X7 & X8 & X9 & X10 \end{array} \\ \left(\begin{array}{cccccccccc} \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & & & & \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & & & & \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & & & & \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & & & & \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & \ominus & \ominus & \ominus & \ominus \\ \otimes & \otimes & \otimes & \otimes & \otimes & \otimes & \ominus & & & \\ & & & & \ominus & \ominus & \oplus & \oplus & \oplus & \oplus \\ & & & & \ominus & & \oplus & \oplus & \oplus & \oplus \\ & & & & \ominus & & \oplus & \oplus & \oplus & \oplus \\ & & & & \ominus & & \oplus & \oplus & \oplus & \oplus \end{array}\right) \end{array}$$



Node	Name
X1	John
X2	Boy
X3	Window
X4	Stare
X5	See
X6	Action[look]
X7	Agent[John]
X8	Glass
X9	Possessor[John]
X10	Instrument[Optical Inst.]

Fig. 10. Corresponding context matrix $\mathbf{M}$ and its labeled graph in scenario a). Context cue Q is activated by the lexical item *telescope* in the current sentence and is blended into the knowledge matrix $\mathbf{K}$ when (S-4) is under analysis.



Node	Name		
1	See[John, with(Girl, Tele.)]	11	Look[John, Girl]
2	with(Girl, Tele.)	12	John
3	Possessor[Girl]	13	Possessor[John]
4	Hold[Girl, Tele.]	14	Telescope
5	Agent[Girl]	15	Glass
6	Look[Girl, John]	16	Optical Instrument
7	Universe	17	See
8	See[with(John, Tele.), Girl]	18	Instrument
9	Agent[John]	19	Hold[John, Tele.]
10	with(John, Tele.)	20	Girl

Fig. 11. Mesh plot of the normalized activation values in a set of 20 linguistic items in the final semantic resolution after the related context cue Q is carried over from the foregoing sentences.

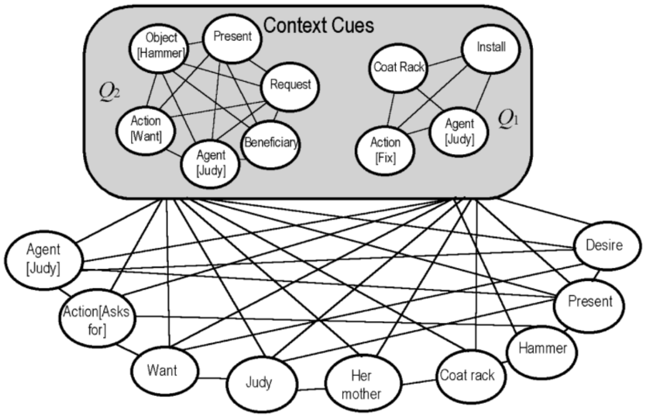


Fig. 12. Knowledge matrix of the sentence. [*She*]$_8$ *asks for* [*her mother*]$_9$ *to get*[*it*]$_{10}$ *for*[*her*]$_{11}$ with the two possible candidates of pronoun *it*.

tence (S-4) and can then be blended together in the semantic resolution process. The resulting activation for the two competing propositions and their coalition members are shown in Fig. 11.

For a better visualization, Fig. 11 shows the normalized activation values in a set of 20 selected linguistic items in the resolution process of sentence (S-4). The three-dimensional mesh plot is interpreted by noting that the height of each point in the mesh plot corresponds to the activation of a linguistic item in the resolution. All points lying on the same horizontal line correspond to the same linguistic item at different points in time. All points lying on the same vertical plane correspond to the activation value of all linguistic items associated with a particular instant of time where time is ordered from the left-hand side of the graph to the right-hand side. Influenced by the closed proximity of *telescope* and *optical instrument* as well as the parallelism of grammatical roles of *Agent*[*John*] and *Possessor*[*John*] from the context cue in the scenario, the system strengthens the corresponding links in the current knowledge matrix and leads to the dominant roles of *Hold*[*John, Telescope*]. As a result, the linguistic item *With*[*John, Telescope*] is boosted and the framework identifies the linguistic item *See*[*With*(*John, Telescope*), *Girl*] to be the best-fit in scenario a).

A more complicated simulation of pronoun resolution is shown in the following excerpt.

> [*Judy*]$_1$ *is going to have* [*a birthday party*]$_2$. [*She*]$_3$ *wants a* [*hammer*]$_4$ *for a* [*present*]$_5$. *Then* [*she*]$_6$ *can fix her* [*coat rack*]$_7$. [*She*]$_8$ *asks for* [*her mother*]$_9$ *to get* [*it*]$_{10}$ *for* [*her*]$_{11}$. [*Her mum*]$_{12}$ *thinks* [*girls*]$_{13}$ *should not play with* [*hammers*]$_{14}$. [*She*]$_{15}$ *buys* [*her*]$_{16}$ [*a dress*]$_{17}$.

This excerpt contains a total of 17 simple substantives where entities 3, 6, 8, 10, 11, 15, and 16 are pronouns. In the upcoming demonstration, we try to resolve the problem of pronoun referents, the entities 10, 15, and 16 as being *a hammer*, *her mum*, and *Judy* respectively. As discussed above, our pronoun resolution is actually composed of pairs, grouping the currently analyzed sentence with the other potential candidates. Resolving an anaphoric reference requires searching the last surviving candidate in the resolution. For each encountered pronoun, it is patently infeasible for all previously introduced candidates to be examined. Since the correct candidate is likely to be fairly close to the pronoun, the number of comparisons is restricted to an arbitrary number in order to reduce the computational complexity. In our experiments, candidates are presented to the resolution in a linear basis, i.e., from the closest to farthest. Fig. 12 shows the knowledge matrix so formed in identifying the entity 10. In order to avoid unnecessary generation of alternatives as most syntactic systems tend to have, we limit our resolution process of entity 10 with the following possible candidates: [*hammer*]$_4$, [*present*]$_5$, and [*coat rack*]$_7$.

The involvement of buffers, which capture the relevant context cues, is designed to carry the preceding analyzed knowledge sources over into the current processing cycle, in the hope that they will serve as common bridging elements between the sentences. As shown in Fig. 12, initial hypotheses are constructed by taking the anaphora *it* and the potential candidates. At the same time, the context cues Q_1 and Q_2 are linked with the corresponding concepts in the sentence. Due to the strong association between the linguistic items *Action*[*Want*], *Action*[*Asks for*], and the co-occurrence of *Agent*[*Judy*] in both sentence and Q_2, the degree of repetition between the current sentence and the context cues allows the system to identify Q_2 to be relevant to the current sentence. The system resolves *it* as *hammer* by virtue of the fact that *hammer* is the object of the *Action*[*Want*] in the context cue. In fact, linguistic studies also show there exists a pervasive form of *linguistic inertia* which manifests as a preference to assign the referent of a pronoun to the linguistic entity in the discourse context that filled the corresponding semantic case role earlier in the text. This is a generalized form of case-role parallelism, which has been proven crucial in anaphora and ellipsis resolution [66]. This simulation also demonstrates our architecture is able to capture linguistic inertia in the semantic interpretation through our context cues in the context generation.

VIII. Evaluations

Our approach using the context-dependent model expresses the meaning conveyed by and the relations among all constituents of the given sentence. Our experiments involving more than 2000 sentences in two different languages (English and Chinese) will be evaluated in this section. Two sets of evaluation are carried out. First, we attempt to gain some indirect measure of the context-dependent models by comparing the degree of coincidence of the context-dependent models produced from a bilingual corpus. It is on the basis that our context-dependent models represent a language-neutral semantic representation which is in strong agreement across different languages. Second, we adopt a statistical approach to analyze the performance of our context-dependent models in pronoun resolution, followed by a summary on how we advance the state-of-the-art.

Our first experiment is based on a bilingual corpus with sentences extracted from two versions of the Bible: the English New International version and the Chinese Union version from the International Bible Society. Because of the diversity in the treatment of Chinese grammar, we have experimented in the evaluation with Chinese sentences that cause only minor structural ambiguities. Most of the sentences describe sequences of events that follow one another in approximately linear temporal sequences. The corpus consists of 313 pairs of parallel verses that have 3908 words and 4235 words in Chinese and English, respectively. In the syntactic parser, more than 30 rules are used. The total number of frames acquired in order to generate the corresponding p-terms in the system is more than 220 with the average number of frames/verb equal to 3.7, as described in Section IV. The average sentence length in Chinese and English are 12.5 and 13.5 words, respectively. In the lexical subsymbols, we have already generated more than 4000 lexical subsymbols in the evaluation. The subsymbols generated cover more than 25% of the Longman lexicon. The context-dependent model for each parallel verse is generated. They are evaluated by assigning a grade manually based on the number of nodes which coincided in the context-dependent models of the parallel verses. One of four grades for the degree of coincidence found in each parallel verse is assigned.

1) Perfect—a perfect match. This is when more than 90% of the linguistic items conveyed from the source languages, in both English and Chinese, coincide in their context-dependent models.
2) Acceptable—all important linguistic items are represented correctly, but some unimportant details are missing. Not less than 70% of coincidence can be found.
3) Poor—the matching between the nodes in the context-dependent models of the parallel verses is greater than 40% but less than 70%
4) Incorrect—unacceptable or the degree of coincidence is found to be too low in both context-dependent models.

The judgements are provided by two or more independent graders. When different judgements exist for a verse, the majority vote is accepted. Moreover, in order to differentiate the longer verses which may generate a relatively complicated representation, we partition the corpus into two main types of verses. Verses with more than 13 tokens will be classified as long verses, and we limit the maximum verse length to be 20 tokens long. Table III shows the percentage of verses under the four different grades.

TABLE III
Subjective Evaluation of the Context-Dependent Models so Formed in a Bilingual Corpus: Percentage of Verses Evaluated as Perfect, Acceptable, Poor, or Incorrect

	Short Verses	Long Verses
Perfect	35%	25.6%
Acceptable	53.2%	44.5%
Poor	8.7%	19.5%
Incorrect	3.1%	10.4%

TABLE IV
Subject and Topic in the Chinese Sentence *Mary her Younger Brother Really Likes Candy* [a]. Both Sentences [b] and [c] in Chinese With Different Word Orders Carry the Same Meaning

[a]	Mary her younger brother really likes candy					
[b]	ma3-li4	ta1	di4-di0	tang2-guo3	zhen1	xi3-huan1
[b]′	ma-li	s/he	younger brother	candy	really	likes
[c]	ma3-li4	ta1	di4-di0	zhen1	xi3-huan1	tang2-guo3
[c]′	ma-li	s/he	younger brother	really	likes	candy

(The subject is (ta1) di4-di0, the topic is ma3-li4 in both sentences [b] & [c])

More than 88 and 70% of the short aligned and the long aligned verses, respectively, in the bilingual corpus are found to have a high degree of coincidence in their context-dependent models. Some characteristics of Chinese can account for why this language-neutral semantic representation is more favorable in language processing. The element *topic* is one of the most striking and prominent features that sets Chinese apart from other languages [67].

What distinguishes *topic* from subject is that the subject must always have a direct semantic relationship with the verb as one that performs the action or exists in the state named by the verb, but the *topic* does not. Table IV illustrates the differences. The example in Table IV describes the primacy of semantic structure in the sentence, while subject-predicate relation emphasizes the tighter relationship of subject-verb-object (SVO) patterns of English. Word order difference between languages is an important consideration in any language processing, such as machine translation. Unlike English, Chinese syntax does not carry as much significance as English does [68]. While the syntactic order of sentence [c] may yield an English parallel, there is no English equivalent for sentence [b]. Given that word order differences between any pair of languages is fairly arbitrary, the use of any rigid grammatical devices, such as phrase structure grammar, poses many challenges in any generation of context representation. On the contrary, our context-dependent models are independent of any surface form of linguistic expression and therefore enable us to bypass many syntactic variations. The context-dependent model is a complex collection of information that contains the *distilled* and *unambiguous* linguistic items. The model captures semantic meaning and can be regarded as the deeper structure underlying the linguistic expressions. As shown in the evaluation, our context-dependent model serves a means to extract deep textual knowledge in both English and Chinese languages.

TABLE V
INFLUENCES OF THE CONTEXT-DEPENDENT MODES IN PRONOUN RESOLUTION

	Recall	Precision	F
Baseline model	51%	36%	42%
Syntactic Network (SN)	38%	31%	34%
SN + Lexical Subsymbols	78%	87%	82%
SN + Lexical Subsymbols + Context-Dependent models	90%	94%	92%

Our next evaluation concentrates on applying the framework on pronoun resolution. Pronouns do not introduce new entities into sentences since they refer to prior entities, as discussed in Section VII. In this evaluation, 518 isolated sentences from the Bible are selected. Each contains third person pronouns (including possessive) or the reflexive pronouns that do not occur very often in the test corpus. All the isolated sentences with more than 600 pronouns are used as the test cases. For each of the isolated test sentences, two preceding sentences are also involved in order to form a coherent portion of text. In other words, 1500 sentences with more than 3000 different lexical items are involved in our pronoun resolution. The average sentence length is 11.8 tokens per sentence. As with other document analysis, the effectiveness of pronoun resolution appears to be dictated by *recall* and *precision* parameters, where recall (R) is a percentage of how many correct pronouns can be identified, while precision (P) is the percentage of pronouns tackled by the system which are actually correct. In addition, a common parameter F is used as a single-figure measure of performance [69].

$$F = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}.$$

We set $\beta = 1$ to give no special preference to either recall or precision. In order to evaluate the effectiveness of the approach, we test our approach with a baseline model. In the model, the most immediate entity, which is introduced into the text preceding the pronoun and agrees with gender and number constraints, is estimated to be the right candidate. Moreover, we show the importance of including all the components of our system by listing the performance of the system when some components are detached. The results obtained are shown in Table V.

As can be seen in the Table V, the drastic increase in F measure is caused by the introduction of lexical subsymbols. The set of possible antecedents of the pronouns tends to be reduced drastically by the linguistic knowledge constraints. This is strongly supported by the case role inertia as well as the lexical subsymbols which promote complementary linguistic primitives of the preferred antecedents for the pronouns. On the other hand, in order to test the robustness of our subsymbols in resolving pronouns, we gradually replace the lexical subsymbols with some random generated bit patterns in order to investigate the sensitivity of our approach. Fig. 13 shows the graceful degradation of the performance of the system.

IX. CONCLUSION

Most of the on-going computational linguistic theories do not define precisely the concept of linguistic context. Many approaches in word-sense disambiguation account for context simply by using windows of surrounding words [70]. This approach is insufficient for discourse theories which need to produce higherlevel context inferences using elaborate information pertaining to the situation conveyed by the foregoing sentences. In this paper, we have focused on the examination of a multifaceted knowledge approach using symbolic and connectionist techniques in learning the contextual effects in the dynamic context generation. We describe a context representation which is learned from a set of diversity linguistic knowledge sources. Our system works incrementally, allows modules running in parallel and integration of a diversity of knowledge sources in the early stage. The approach is on the basis that two word-senses occurring in the same sentence will probably be semantically related. This prediction bootstraps an in-depth processing in context generation. Moreover, there are several other reasons which make our system advance the state-of-the-art in natural language processing.

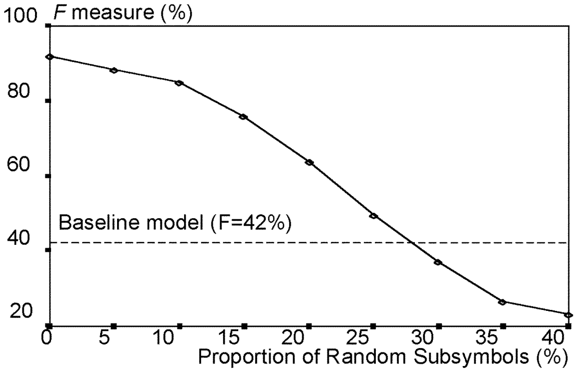


Fig. 13. Sensitivity analysis of the performance while the proportion of random generated subsymbols varies.

In most semantic interpretation, while it is recognized that context makes a great contribution to understanding, the mechanism of how the context is identified and utilized is not seriously addressed. Our approach concentrates on the context cue formation from the standpoint of the major linkages among sentences. The context-dependent model captures the information which has already been activated in the discourse focus dynamically. The model not only provides a means to represent the distilled information from the previously analyzed sentences, but it prescribes an effective mechanism for combining sentential knowledge which can be carried over from the preceding sentence into the current processing cycle. The model is involved when the current sentence interacts with the foregoing analyzed utterance in the following ways:

1) by strengthening an existing hypothesis;
2) by contradicting and eliminating an existing assumption;
3) by combining with an existing assumption to yield a contextual implication, i.e., a logical implication derivable neither from the current sentence, nor from the context alone, but from the current sentence and the context combined.

While syntactically based theories of language understanding inevitably focus on individual sentences, our approach suggests that there is more to integrating the linguistic information from the different sources than juxtaposition. We start out from the syntax and case structures to facilitate the syntactic network of p-terms. The syntactic component converts phrasal patterns into p-terms expressing the compositional and recursive structures of the language. Each p-term is supported by a set of lexical subsymbols. The word associations in our subsymbols are strong sources of information that a reader must weigh against other cues since they make immediate and obvious sense selections. While the syntactic network provides the depth of sentential meaning decomposition, the lexical symbols provide the bredth of the possible lexical meaning. On the other hand, close analysis of the context cues generated from sentences reveals that a deep structure, which is independent of any surface form of linguistic expressions captures basic semantic meaning. These context cues, establishing the links between sentences, are gradually integrated into the understanding process.

The resolution mechanism consists of inferences from multiple linguistic information sources. It copes with several competing concept hypotheses and aims at a constraint satisfaction-based selection among the linguistic primitives. The right candidate for the semantic resolution receives a sufficient amount of positive activation from multiple linguistic information sources and low inhibitory activation from the others. The system captures the right candidates under the constraint satisfaction mechanism. Our context models are constructed incrementally by incorporating distilled information both explicitly and implicitly stated in the foregoing sentences.

Unlike some other approaches to tasks such as anaphora resolution, our approach does not regard it as a separate issue from language understanding, but treats it simultaneously with the semantic interpretation. The knowledge that serves as a basis for the system reasoning includes information about language and its use. Lexical associations, syntactic restrictions, case-role expectations, as well as contextual effects are represented and utilized not only in anaphora resolution but also for language understanding. The contribution of each knowledge source is not interpreted as a scoring procedure that provides a confidence measure as used in other systems. Instead, to support robust resolution, a framework allowing for different linguistic knowledge sources, although diverse in their natures, is devised. The independent linguistic knowledge sources, in the framework, can be seen as remarkably compatible with each other—complementary rather than in conflict. Our semantic resolution process is concerned with not just the question of finding the appropriate linguistic items required but the matter of suppressing those that are irrelevant. The process is to narrow down from disparate sources to the most plausible referent and search for the last surviving candidate that will make the sentences a coherent whole.

Although it is necessary to oversimplify the reality in the experiments so as to get a system that we could easily work with, the resulting system and the simulations demonstrated are sufficiently rich to have a realistic structure. Our experiments have shown some promising results in some domains of language processing. However, taking up the full challenge toward a more powerful, unrestricted and linguistic-oriented framework is certainly not without pains. First, features for semantic representations of the lexical items are based on the semantic hierarchy associated with each word from the lexicon and then augmented by hand. Although the lexicon includes entries for many general class items, features for some domain-specific lexical items in the training sets have to be handcrafted. On the other hand, the decomposition of the analyzed sentences into p-terms from a set of case frame structures and the organization of the p-terms all require minimum manual intervention. We expect it will take more efforts in scaling up the whole system in unrestricted domains. In fact, the parser is currently being extended into a case-based parser that will produce a crude conceptual representation. Refining this representation will be preceded by more work on the parser to separate conjoined clauses and phrases into fragments directly corresponding to objects and activities. Clearly, it is too ambitious to claim that our system can handle all the problems in language understanding. Additional experimental work is certainly required; nevertheless, the simulations have presented its capabilities in utilizing context in natural language understanding. In short, our model does not claim to deal with all aspects of language, but its limitations are not relevant to our main focus: Our context dependent representation can be learnt from grammatical and even ungrammatical sentences using much simpler and more local types of linguistic primitives without demanding a complete syntactic analysis.

In conclusion, the field of natural language understanding has witnessed an unprecedented surge of interest in empirical methods. A recent trend has been to try to circumvent many of the syntactic and semantic complexities of natural language by aiming to extract only certain predetermined sets of information from narrowly focused classes of texts. While such an approach can achieve high quantitative returns, it necessarily compromises the quality of understanding. In this paper, we advocate achieving deeper understanding as an important and realistic goal in the community. The main concern of our research has been to develop an adequate context-dependent representation for language understanding systems, especially ones aimed at understanding narratives. The representation not only provides ease of access to multifaceted linguistic knowledge, but is able to utilize it effectively. Our dynamically context-based language understanding gives a method of a judicious resolution of the collection of evidences which emerge from probabilistic inferences for the surface syntactic structure, semantic association in our lexical subsymbols as well as learning algorithms in capturing the enriched context cues. A concrete framework with experimental results is demonstrated. The evaluations lead to two major findings. First, with the incorporation of more explicit lexical subsymbols and contextual information, the overall performance will be boosted in language understanding, such as in anaphora resolution. Second, the high degree of coincidence between the context-dependent models of English and Chinese illustrates the usefulness of the language neutral semantic representation. Our approach shows promise for enhancing the robustness and accuracy of language understanding.

References

[1] S. Pinker, *Language Learnability and Language Developments*. Cambridge, MA: Harvard Univ. Press, 1984.

[2] C. Cherry, *On Human Communication*. Cambridge, MA: MIT Press, 1975.

[3] A. N. Chomsky, *Syntactic Structures*. Paris, France: Mouton, 1957.

[4] M. A. Gernsbacher, *Language Comprehension as Structure Building*. Hillsdale, NJ: Lawrence Erlbaum, 1990.

[5] M. A. K. Halliday and C. M. Matthiessen, *Construing Experience Through Meaning*. New York: Cassell, 1999.

[6] J. Pustejovsky and B. Boguraev, "Lexical knowledge representation and natural language processing," *Artif. Intell.*, vol. 63, no. 1–2, pp. 193–223, 1993.

[7] P. A. Carpenter and M. Daneman, "Lexical retrieval and error recovery in reading: A model based on eye fixations," *J. Verbal Learn. Verbal Behav.*, vol. 20, pp. 137–160, 1981.

[8] W. Kintsch, "A cognitive architecture for comprehension," in *Cognition: Conceptual and Methodological Issues*, H. L. Pick, P. van den Broek, and D. C. Knill, Eds. Washington, DC: Psychol. Assoc., 1992, pp. 143–163.

[9] C. R. Fletcher and C. P. Bloom, "Causal reasoning in the comprehension of simple narrative texts," *J. Mem. Lang.*, vol. 27, pp. 235–244, 1988.

[10] S. C. Shapiro, "SNePS: A logic for natural language understanding and commonsense reasoning," in *Natural Language Processing and Knowledge Representation*, L. M. Iwanska and S. C. Shapiro, Eds. Menlo Park, CA: AAAI, 2000, pp. 175–195.

[11] J. R. Hobbs, M. E. Stickel, D. E. Appelt, and P. Martin, "Interpretation as abduction," *Artif. Intell.*, vol. 63, pp. 69–142, 1993.

[12] E. Charniak and R. Goldman, "A semantics for probabilistic quantifier-free first-order languages, with particular application to story understanding," in *Proc. 11th Int. Joint Conf. Artificial Intelligence*, vol. 2, 1989, pp. 1074–1079.

[13] Y. Bar-Hillel, "Automatic translation of languages," in *Advances in Computers*, D. Booth and R. E. Meagher, Eds. New York: Academic, 1960, pp. 23–37.

[14] G. W. Cottrell and S. L. Small, "A connectionist scheme for modeling word sense disambiguation," *Cogn. Brain Theory*, vol. 6, pp. 89–120, 1983.

[15] D. L. Waltz and J. B. Pollack, "Massively parallel parsing: A strongly interactive model of natural language interpretation," *Cogn. Sci.*, vol. 9, pp. 51–74, 1985.

[16] T. E. Lange, "Lexical and pragmatic disambiguation and reinterpretation in connectionist networks," *Int. J. Man-Mach. Stud.*, vol. 36, pp. 191–220, 1992.

[17] M. F. St. John and J. L. McClelland, "Learning and applying contextual constraints in sentence comprehension," *Artif. Intell.*, vol. 46, pp. 217–257, 1990.

[18] N. Sharkey, *Connectionist Natural Language Processing*. Norwell, MA: Kluwer, 1992.

[19] R. Miikkulainen, *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press, 1993.

[20] E. Brill, "Some advances in transformation-based part of speech tagging," in *Proc. 12th Nat. Conf. Artificial Intelligence*, 1994, vol. 1, pp. 722–727.

[21] D. Yarowsky, "One sense per collocation," in *Proc. ARPA Human Language Technology Workshops*, NJ: Princeton, 1993, pp. 266–271.

[22] B. Merialdo, "Tagging English text with a probabilistic model," *Comput. Linguist.*, vol. 20, no. 2, pp. 155–171, 1994.

[23] L. M. Iwanska and S. C. Shapiro, *Natural Language Processing and Knowledge Representation*. Menlo Park, CA: AAAI, 2000.

[24] C. Cardie, "Empirical methods in information extraction," *AI Mag.*, vol. 18, no. 4, pp. 65–80, 1997.

[25] R. Basili, M. T. Pazienza, and P. Velardi, "An empirical symbolic approach to natural language processing," *Artif. Intell.*, vol. 85, pp. 59–99, 1996.

[26] G. A. Miller, "WordNet—A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[27] S. W. McRoy, "Using multiple knowledge sources for word sense discrimination," *Comput. LinguisT.*, vol. 18, no. 1, pp. 1–30, 1992.

[28] L. A. Urena, J. M. G. Hidalgo, and M. de-Buenaga, "Information retrieval by means of word sense disambiguation," in *Lecture Notes in Artificial Intelligence*. New York: Springer-Verlag, 2000, vol. 1902, pp. 93–98.

[29] S. M. Harabagiu and D. I. Moldovan, "A parallel system for text inference using marker propagations," *IEEE Trans. Parallel Distrib. Syst.*, vol. 9, pp. 729–747, Aug. 1998.

[30] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, vol. 1, 1995, pp. 448–452.

[31] C. Leacock, G. Towell, and E. M. Voorhees, "Toward building contextual representations of word senses using statistical methods," in *Corpus Processing for Lexical Acquisition*, B. Boguraev and J. Pustejovsky, Eds. Cambridge, MA: MIT Press, 1996, pp. 97–113.

[32] S. W. McRoy, "Achieving robust human-computer communication," *J. Human-Comput. Studies*, vol. 48, no. 5, pp. 681–704, 1998.

[33] M. Redington, N. Chater, and S. Finch, "Distributional information: A powerful cue for acquiring syntactic categories," *Cogn. Sci.*, vol. 22, no. 4, pp. 425–469, 1998.

[34] G. Hirst, *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[35] W. Gale, K. Church, and D. Yarowsky, "One sense per discourse," in *Proc. DARPA Speech Natural Language Workshop*, New York, 1992, pp. 233–237.

[36] D. A. Cruse, *Lexical Semantics*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[37] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Mach. Learn.*, vol. 7, pp. 195–225, 1991.

[38] T. McArthur, *Longman Lexicon of Contemporary English (English-Chinese Edition)*. White Plains, NY: Longman, 1992.

[39] Chinese Knowledge Information Processing Group, *Academia Sinica Balanced Corpus, Version 3.0*. Taipei, Taiwan, R.O.C.: Inst. Inform. Sci., Academia Sinica, 1998.

[40] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Inf. Process. Manage.*, vol. 33, pp. 193–207, 1997.

[41] G. E. Hinton, "Representing part-whole hierarchies in connectionist networks," Connectionist Res. Group, Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG TR-882, 1988.

[42] J. B. Pollack, "Recursive distributed representations," *Artif. Intell.*, vol. 46, pp. 77–105, 1990.

[43] P. Smolensky, "Tensor product variable binding and the representation of symbolic structures in connectionist systems," *Artif. Intell.*, vol. 46, pp. 159–216, 1990.

[44] P. Gupta and D. S. Touretzky, "Connectionist models and linguistic theory: Investigations of stress systems in language," *Cogn. Sci.*, vol. 18, pp. 1–50, 1994.

[45] A. Sperduti, "Stability properties of labeling recursive auto-associative memory," *IEEE Trans. Neural Netw.*, vol. 6, pp. 1452–1460, Nov. 1995.

[46] T. A. Plate, "Holographic reduced representations," *IEEE Trans. Neural Networks*, vol. 6, pp. 623–641, May 1995.

[47] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1 and 2.

[48] D. S. Blank, L. A. Meeden, and J. B. Marshall, "Exploring the symbolic/subsymbolic continuum: A case study of RAAM," in *Closing the Gap: Symbolism vs. Connectionism*, J. Dinsmore, Ed. New York: Lawrence Erlbaum, 1992.

[49] M. G. Dyer, M. Flowers, and Y. J. A. Wang, "Distributed symbol discovery through symbol recirculation: Toward natural language processing in a distributed connectionist network," in *Connectionist Approaches to Natural Language Processing*, R. G. Reilly and N. E. Sharkey, Eds. New York: Lawrence Erlbaum, 1992, pp. 21–48.

[50] Y. Wilks, D. Fass, C.-M. Guo, J. McDonald, T. Plate, and B. Slator, "Providing machine tractable dictionary tools," in *Semantics and the Lexicon*, J. Pustejovsky, Ed. Norwell, MA: Kluwer, 1993, pp. 341–401.

[51] J. A. Fodor and Z. W. Pylyshyn, "Connectionism and cognitive architecture: A critical analysis," in *Connections and Symbols*, S. Pinker and J. Mehler, Eds. Cambridge, MA: MIT Press, 1988, pp. 3–71.

[52] C. J. Fillmore, "The case for case," in *Universals in Linguistic Theory*, E. Bach and R. T. Harms, Eds. New York: Holt, Rinehart & Winston, 1968, pp. 1–90.

[53] T. V. Geetha and R. K. Subramanian, "Representing natural language with Prolog," *IEEE Softw.*, vol. 7, pp. 85–92, Mar. 1990.

[54] R. F. Simmons, "Semantic networks: Their computation and use for understanding English sentences," in *Computer Models of Thought and Language*, R. C. Schank and K. M. Colby, Eds. San Francisco, CA: Freeman , 1973, pp. 63–113.

[55] J. Pearl, *Probabilistic Inference in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann, 1988.

[56] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*. Hillsdale, NJ: Lawrence Erlbaum, 1998.

[57] S. Stoddard, *Text and Texture: Patterns of Cohesion, Advances in Discourse Processes*. Norwood, NJ: ABLEX, 1991, vol. XL.

[58] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[59] M. Hoey, *Patterns of Lexis in Text*. Oxford, U.K.: Oxford Univ. Press, 1991.

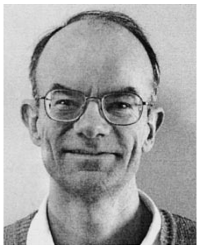
[60] H. Jackson, *Words and Their Meaning*. London, U.K.: Longman, 1988.

[61] R. Sun, "An efficient feature-based connectionist inheritance scheme," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, pp. 1–12, Mar./Apr. 1993.

[62] J. A. Hendler, "Marker-passing over microfeatures: Toward a hybrid symbolic/connectionist model," *Cogn. Sci.*, vol. 13, pp. 79–106, 1989.
[63] L. A. Bookman, *Trajectories Through Knowledge Space: A Dynamic Framework for Machine Comprehension*. Norwell, MA: Kluwer, 1994.
[64] S. W. K. Chan and J. Franklin, "Symbolic connectionism in natural language disambiguation," *IEEE Trans. Neural Networks*, vol. 9, pp. 739–755, Sept. 1998.
[65] S. W. K. Chan, "Integrating linguistic primitives in learning context-dependent representation," *IEEE Trans. Knowledge Data Eng.*, vol. 13, pp. 157–175, Mar./Apr. 2001.
[66] J. G. Carbonell and R. D. Brown, "Anaphora resolution: A multi-strategy approach," in *Proc. Int. Conf. Computat. Linguist.*, 1988, pp. 96–101.
[67] C. N. Li and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles, CA: Univ. of California Press, 1981.
[68] C.-R. Huang and K.-J. Chen, "Issues and topics in Chinese natural language processing," in *Readings in Chinese Natural Language Processing*, C.-R. Huang, K.-J. Chen, and B. K. T'sou, Eds., 1996, Monograph No. 9, pp. 1–22.
[69] C. J. van Rijsbergen, *Information Retrieval*, Second ed. London, U.K.: Butterworths, 1979.
[70] Y. Wilks and M. Stevenson, "Combining independent knowledge sources for word sense disambiguation," in *Proc. Int. Conf. Recent Advances in Natural Language Processing*, Tsigov Chark, Bulgaria, 1997, pp. 1–7.

Samuel W. K. Chan (A'96) received the M.Sc. degree from the University of Manchester, Manchester, U.K. in 1986, the M.Phil. degree from the Chinese University of Hong Kong (CUHK), in 1991, and the Ph.D. degree from the University of New South Wales, Sydney, Australia in 1998, all in computer science.

He is currently an assistant professor with the CUHK. His research interests include applying machine learning techniques in computational linguistics, content-based information retrieval, and data mining with emphasis on text.



James Franklin received the Ph.D. degree in algebra from Warwick University, Warwick, U.K., in 1981.

He is the author of *The Science of Conjecture: Evidence and Probability Before Pascal* (Baltimore, MD: Johns Hopkins Univ. Press, 2001) and *Introduction to Proofs in Mathematics* (New York: Prentice-Hall, 1988). His research interests include the philosophy of mathematics and the use of cluster analysis in symbol grounding.