Exploring the generation and usage of inference in a biologically inspired neural network system

Eleni Koutsomitopoulou

Georgetown University, Washington DC, USA
LexisNexis UK
Eleni.Koutsomitopoulou@lexisnexis.com

Abstract

Inference allows speakers to extract knowledge about facts that are not explicitly existent in a document or discourse. In particular, it allows for drawing new (i.e. previously unknown and/or not explicitly stated) conclusions about the relationships of known and stated facts (topics, events, entities) of the (NL) language input. Logical deduction and induction –and hence learning– are based on this inferential process.

This paper sketches some aspects of a neural network algorithm pertinent to inference generation and usage. The *Adaptive Resonance Theory Zero* (ART0) neural network system is not designed to model inference with explicit rules. Instead, it represents the knowledge the reader is expected to draw from the facts stated in the input, and based on this representation it makes inferences about certain implied facts and their relationships. The basic hypothesis tested with this model is that the short-term (STM) and long-term memory (LTM) systems play a crucial role in learning, knowledge representation and by extension in inferencing. A ramification of this hyposthesis is that a model of the short-term and long-term memory systems is in fact a model suitable for inference.

1 Introduction

Symbolic inference is defined as the "deduction of new (i.e. previously unknown) facts using existing facts" [1]. Question answering systems are the most obvious usages of systems capable of performing inference. Logical deduction and induction –and hence learning– are based on this inferential process. In particular, performing inference allows speakers to extract knowledge about facts that are not existent in a document. Unlike basic extraction systems, inference machines are capable of drawing previously unknown conclusions about the relationships of known and stated facts (topics, events, entities) of the natural language (NL) input.

The ART0 neural network system is not designed to model inference with explicit rules. Instead it represents the knowledge the reader is expected to draw from the facts stated in the input and based on this representation it makes inferences about certain implied facts and their relationships. The basic hypothesis tested with this model is that the short-term (STM) and long-term memory (LTM) systems play a crucial role in learning, knowledge representation and by extension in inferencing.

The inferential process is guided by variable binding in the sense that "unknown" (i.e. unidentified or ambiguous) factoids of the NL input act as variables that they get bound to a particular stated elements of the argument structure of the propositional input that has been previously learnt by the system. This binding process facilitates inference, question answering and learning. In vivo, neuronal activation patterns replace the need for variable binding in NL introducing essentially a variable-free grammar.

This paper sketches some aspects of the ART0 system pertinent to inference generation and usage.

2 About the ART0 network

The ART0 system is a neural network simulation capable of demonstrating how neuronal activation influences the shape of the patterns that particular linguistic constructs form when they are uttered and understood by speakers of a language. Two types of memory systems are maintained: STM and LTM. The STM system simulates in vivo activation patterns of NL input at time tx. The LTM system stores the weighted connections between nodes in the network. Each node represents a linguistic terminal element extracted from the parse tree that is the output of the parser applied to the original NL input to the system.

Knowledge of different domains ("discourses") is "encoded" in the form of pertinent NL input. The basic premise here is that each discourse (or domain) is defined by what speakers know about the main topic in it. There are no a priori defined and assumed sets of facts. Of course natural discourse is inherently elliptical and many of the necessary for inference factoids might not be readily explicit in the chosen piece of discourse. In such cases, I assume that these factoids have been learned at a previous time as part of the natural development and augmentation of the network. For demonstration purposes I model particular examples of coherent natural language discourse that each includes a complete set of factoids that are then used to infer unknown facts, or draw conclusions.

Another critical methodological point of the ART0 system is that no rules apply directly. Instead, I assume that the learning and inference process involves mapping from one domain/discourse to another. There are no "conversion rules" that explicitly map propositions from one domain to another. ART0 maps the output of the parser as nodes in the network and then cal-

culates both the node values and the weights in the connections using two fundamental ART (Adaptive Resonance Theory, [2], [3], [4]) equations one for the STM and the other for the LTM system respectively.

Understanding of an utterance (or piece of text) proceeds by mapping the argument structure (i.e. information about "who did what to whom") of the proposition into the network and let each terminal node in the parse-tree become a node in the network. Each discourse makes a resonant network where all nodes within a sentence excite each other and all sentence-nodes inhibit either other. As two discourses are activated together, certain nodes act like variables and the network needs to decide on their values i.e. binding them to other nodes in one of the two (or more) discourses. This is also a basic case of sense disambiguation, i.e. determining the contextually appropriate meanings of an ambiguous term. Since variable binding is essential for inference and disambiguating variables is part of binding variables to domain-specific values, a system that addresses ambiguity efficiently is expected to fare well on drawing inferences as well. Unlike rule-based systems, ART0: 1) defines its "reasoning space" given the existent factoids that are extracted directly from actual NL input, 2) finds implications of the extracted factoids by extracting the argument structure of the propositions in this discourse, and mapping them in the network, 3) defining the way the mapped factoids interact within two minimally differing discourses by way of node connectivity patterns, 4) reduces the derived implications to specific node activation patterns that are generated using the ART equations for learning and memory. The heart of the system lies in lateral inhibition (mutual inhibition between neighboring neurons/nodes) polarized around minimal dipole anatomies [1]. The inhibitory elements are unbound variables that map into particular terminal nodes of the parse tree as they enter the network. Previous systems taking into account inhibition in NL do not rely on the particular biologically plausible ART equations for Hebbian learning [2] and memory that ART0 uses:

(1) $\dot{x}_j = -Ax_j + Bx_i z_{ij} - Cx_k + I$

(2) $\dot{z}_{ij} = -Dz_{ij} + Ex_i x_j$

In (1) the change of the value of node x_j in time is being calculated based on parameters A, B, C and I. Parameter A is a negative parameter corresponding to the natural decay of the x_j value in time (for instance when there is no excitation or $B = 0$). The parameter B is the learning rate of node x_j. z_{ij} is the change in the weighted connection between node at site x_j and its

[1] Like the linguistic "minimal pairs", minimal dipole anatomies are minimal pairs of mutually inhibitory neurons, which are crucial in the learning process as described by the ART0 algorithm.

[2] "Hebbian learning" is the learning process as described by Hebb D. (1949, *The organization of behavior*. New York: Wiley) that is as the physiological association of neuron A and its neighboring neuron B when A repeatedly causes to fire B. Hebb explained the significance of LTM in retaining the information learned in simple neuronal structures when they are switched off.

excitatory x_i (Hebbian learning). Parameter C is the inhibition rate that the particular x_j node receives from node at site x_k. And parameter I is a form of exogenous input to the x_j sub-network that works in a regulatory way in order to prevent the network general activation from becoming too low or too high. In (2), parameter D is the natural decay at a LTM level of z_{ij} connection. Parameter E is LTM learning rate and it is a function of both the node x_j and its excitatory counter-node.

3 Illustrative example

For illustration, in what follows I present a case of contextual coreference and how it is resolved by the ART0 network.

3.1 Contextual Coreference

3.1.1 Statement of the problem and initial hypothesis

The following is a case of coreference hard to resolve by means of a traditional parser. Paragraphs (A) and (B) below are successive in a coherent document:

> Paragraph A: A witness in the trial of a Moroccan man charged with aiding the Hamburg al-Qaida cell recanted statements to police that he had seen two alleged cell members in Afghan training camps.
>
> Paragraph B: Bekim Adeni on Wednesday threw into doubt an important part of the case against Mounir el Motassadeq, who is charged with belonging to a terrorist organization and with 3,000 counts of being an accessory to murder in the Sept. 11 attacks.

There are two coreferring pairs of nominals in the above two paragraphs: a) the pair *witness−Bekim Adeni* and b) the pair *suspect−Mounir el Motassadeq*. Notice that with the absence of pronouns with known gender and number features coreference is hard to resolve.

This type of coreference, "definite NP coreference", differs from typical coreference phenomena in that it does not involve a pronoun referring to the same entity as a corresponding NP. Instead, in the above two paragraphs, two different NPs corefer to the same entity [3]. In addition, the document contains two pairs of coreferring NPs. The problem is dual: 1) how does the reader disambiguate this particular case of coreference? and 2) how can a biologically inspired algorithm effectively model this process? My hypothesis is that the network is able to learn to: 1) identify the entity to which each pair of NPs are referring and hence 2) distinguish between the referents of the two pairs of NPs.

[3] This type of coreference is atypical not only because of the lack of overt pronouns to anchor the coreference, but they are atypical also because the proper names involved, albeit anaphoric, provide no explicit (feature) information about their coreference ties with the referent common NPs.

3.1.2 Experimental Procedure

In real-time discourse paragraphs A and B are presented successively and both pairs of NPs are simultaneously identified and disambiguated in the discourse. For the purposes of representation, the relevant factoids are analyzed next.

Specifically, for the identification of the *witness*, the following factoids are learnt from paragraphs A and B:

(i) A witness recants statements.

(ii) Recanted statements weaken a case [4]

(iii) A case is against a suspect.

(iv) Bekim Adeni threw into doubt an important part of the case against the suspect [5]

In uttering (1) to (3) at time t the network (like a reader of the document above) learns the factoids depicted by the corresponding propositions. Subsequently, when uttering (4) as phasic input [6] at time $t+1$ the network is presented with new, or marked information about the subject of the recantations. The introduction of this type of phasic input to the network at this point of the learning process results in successfully identifying the referent of the Proper name phrase as *Bekim Adeni*.

Similarly, the factoids below help the reader identify the referent of the NP *a Moroccan man*:

(i) A suspect is charged with a crime.

(ii) Belonging to a terrorist organization is a crime.

(iii) A case is about a crime.

(iv) A case is against a suspect.

(v) A Moroccan man is a suspect.

(vi) Bekim Adeni threw into doubt an important part of the case against Mounir el Motassadeq.

In reading phasic input S6, the reader already knowing factoids S1 to S5, identifies the referent of the NP *a Moroccan man* as well as the referent of the Proper name phrase *Mounir el Motassadeq*, therefore understanding the coreference. Understanding both coreference pairs also prevents erroneous

[4] This is a prelearned factoid. This kind of a priori "word knowledge" makes symbolic learning systems hard to implement. The ART0 system learns from parsed sentences, and disambiguation is obtained given sufficient prelearned relevant factoids mapped on the network.

[5] A more simplified prelearned factoid would be "Bekim Adeni undermined the case against the suspect."

[6] Phasic input is input presented to the network at a later time than a pre-decided learning period. For instance, if all sentences have been introduced and learned at time $t+1$, the phasic input is introduced at the next timestep and learned after every other sentence has already been learned. This way we can attest the effects of learning the particular input.

readings of the otherwise vague references of NPs in paragraphs A and B. Additionally, via the dipole between *witness* and *suspect*, the two coreference pairs are identified and disambiguated simultaneously.

3.1.3 Findings

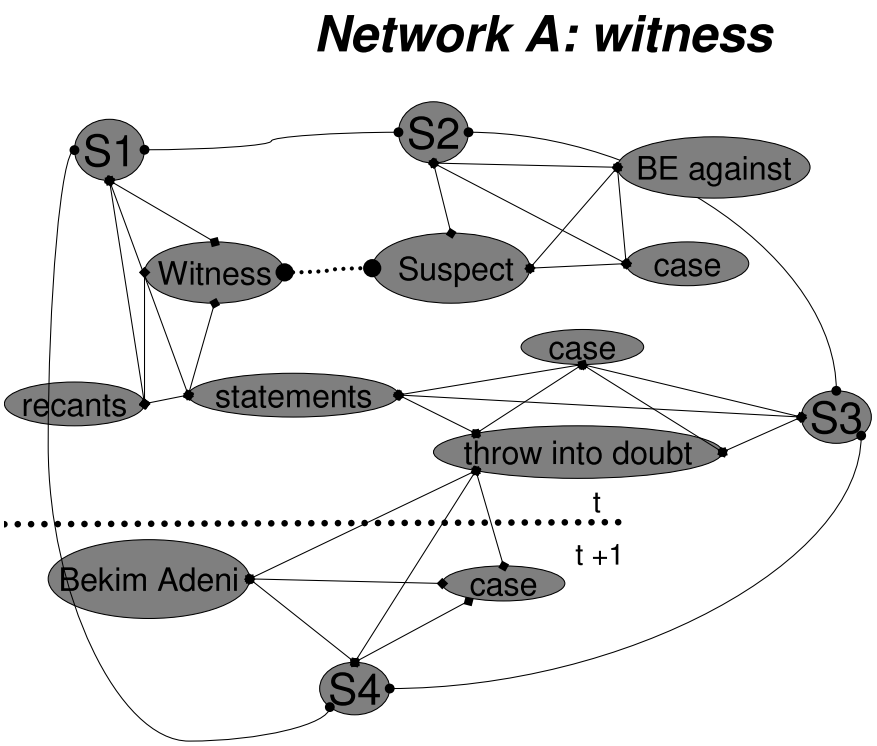

Fig. 1. Network A: Coreference for the *witness* node

Network A consisted of the sentences below:

S1: A witness recants statements.

S2: Recanted statements weaken a case.

S3: A case is against a suspect.

S4: Bekim Adeni weakened a case.

S4 was introduced as phasic input after S1-S2-S3 had been learned. *Witness* and *Suspect* are laterally inhibitory nodes in the network.

Network B consisted of the sentences below:

S1: Bekim Adeni is a witness.

S2: A suspect is charged with a crime.

S3: Aiding a terrorist organization is a crime.

S4: A Moroccan man is aiding a terrorist organization.

S5: Bekim Adeni threw into doubt the case against Mounir el Motassadeq.

In network B, S5 is phasic input and again *Witness* and *Suspect* are mutually inhibitory nodes.

Each network learns a different set of factoids as presented by the discourse.

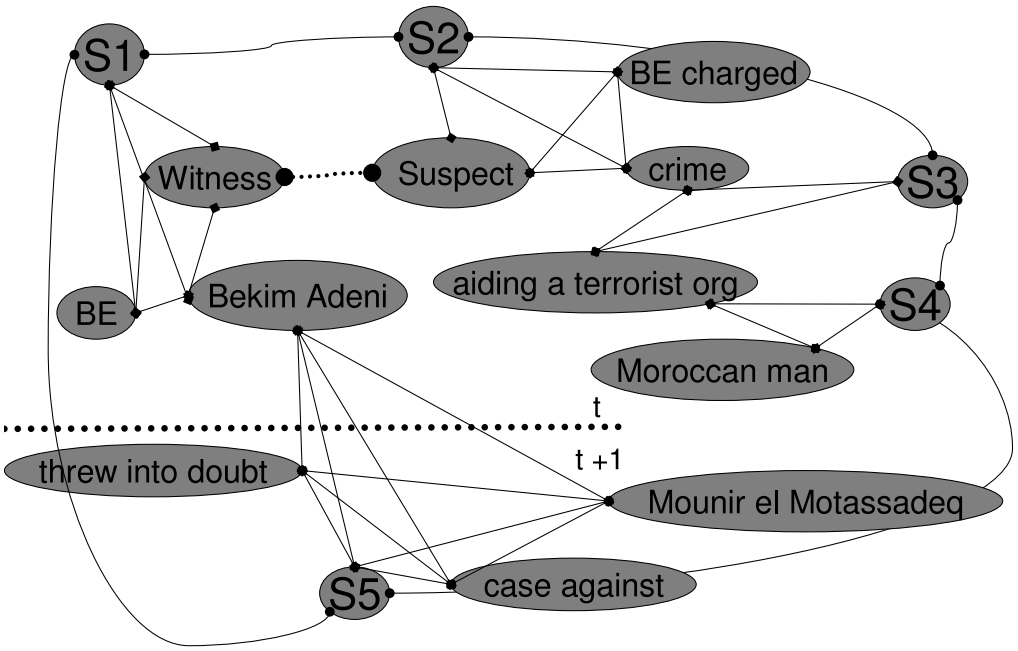

Fig. 2. Network B: Coreference for the *suspect* node

Network A after the introduction of the phasic input learns to distinguish between a suspect vs. a witness, whereas Network B demonstrates how the NP *a Moroccan man* is disambiguated.

The tables below show the pertinent results in terms of x_j node values. In each network, one pole of the inhibitory $witness - suspect$ dipole is expected to be learned. The nodes are learned in the order the corresponding sentences are presented to the network. Each table shows the x_j values for each pole of the dipole for each network at three different timesteps during the learning cycle: t_{S_x}, the time during which all sentences except for the phasic input have been introduced and learned, $t_{S_{x+1}}$, the timestep during which the phsaic input has been introduced and learned, and t_{Stab} or stabilization time, the timestep during which the network has completed learning the entire set of sentences and has achieved "resonance" i.e. a state of mutual excitation and amplification of the signal for learning.

Notice that the x_j value discrepancy for the two nodes is not too significant when paremeter B (learning rate) is .45 as in the experiment for network A. Increasing the value of parameter B to .6 was necessary to get better value discrepancy for the same inhibition rate (parameter $C = .6$). The reason for this required alteration in the learning rate is easily explicable by the fact that network B consists of a larger number of sentences and hence nodes. It is expected that the learning rate needs to be higher proportionately to the number of nodes the network holds.

Node	Time		
	t_{S_x}	$t_{S_{x+1}}$	t_{Stab}
$witness_{S_1}$	9.218	8.470	**8.679**
$suspect_{S_2}$	9.212	8.862	8.082

Parameter	Value
A	0.15
B	0.45
C	0.6
D	
E	
Z_{ij}	0.5 (stable)

Table 1
Phasic input S4 in net A causes amplification of the activation of the node *Witness* in S1

Node	Time		
	t_{S_x}	$t_{S_{x+1}}$	t_{Stab}
$witness_{S_1}$	7.235	9.632	6.657
$suspect_{S_2}$	7.506	9.930	**7.471**

Parameter	Value
A	0.15
B	0.6
C	0.6
D	
E	
Z_{ij}	0.5 (stable)

Table 2
Phasic input S5 in net B causes amplification of the activation of the node *Suspect* in S2

Note that the above two tests only show the STM effects. When LTM is also calculated, parameter B is overshadowed by parameter E. In similar experiments[7], the calculation of the LTM causes the network to yield results with better discrepancy between the inhibitory nodes.

4 Findings

The findings support our initial hypothesis that discourse-level NP-coreference phenomena are accurately represented in and adequately disambiguated by a ART0 network.

In addition via appropriate neuronal activation patterns the ART0 algorithm performs the appropriate binding of anaphoric variables and is proved to be in the right direction as far as inferential reasoning is concerned.

5 Future Work

Further research in both the ART0 system optimisation as well as its large-scale application testing is underway.

References

[1] Sinclair, I. R., *Collins Dictionary of Computing* (2000).

[2] Loritz, D., *How the Brain Evolved language.*, Oxford University Press (1999).

[3] Grossberg, S., *A neural theory of punishment and avoidance. II: quantitative theory.*, Mathematical Biosciences **15** (1972), pp. 253-285.

[4] Grossberg, S., *The adaptive self-organization of serial order in behavior: speech, language, and motor control.*, In Schwab, E.C. and Nusbaum, H.C., Pattern Recognition by Humans and Machines. Orlando: Academic Press, (1986), pp. 187-294.

[7] The coreference case presented here is only one of the several case-studies examined in detail in the author's doctoral dissertation.