Towards a Biologically Inspired Question-Answering Neural Architecture

Derek Monner [a,1] James A. Reggia [a,b]
[a] *Department of Computer Science, University of Maryland, College Park, USA*
[b] *Institute for Advanced Computer Studies, University of Maryland, College Park, USA*

Abstract. Though question-answering systems like IBM's Watson are undoubtedly impressive, their errors are often baffling and inscrutable to onlookers, suggesting that the strategies they use are far different than those that humans employ. Desiring a more biologically inspired approach, we investigate the extent to which a neural network can develop a functional grasp of language by observing question/answer pairs. We present a neural network model that takes questions, as speech-sound sequences, about a visual environment, and learns to answer them with grounded predicate-based meanings. The model must learn to 1) segment morphemes, words, and phrases from the speech stream, 2) map the intended referents from the speech signal onto objects in the environment, 3) comprehend simple questions, recognizing what information the question is asking for, and 4) find and supply that information. Model evaluations suggest that the grounding and question-answering parts of the problem are significantly more demanding than interpreting the speech input.

Keywords. question answering, grounded language comprehension, recurrent neural network, long short term memory

1. Introduction

While question-answering systems such as IBM's recent *Jeopardy!* winner Watson [1] have been well-studied in natural language processing domains, little research has been done into to how the question/answer style of interaction might influence the way humans acquire language. This is an interesting question in light of the fact that, when listening to language, learners are constantly confronted with request/response, question/answer pairs. In this paper we investigate the extent to which a pure neural network model of a human learner can learn a micro-language by listening to question/answer pairs. The model is situated in a simulated micro-world along with two speakers whom we will call Watson and Sherlock. Watson asks questions about the shared environment in a subset of English, and Sherlock responds to these questions with the information Watson was seeking. The model's task is to learn to emulate Sherlock. To do this effectively, the model must listen to the speech sounds of Watson's questions and learn to segment them into morphemes, words, and phrases, which it must then interpret with respect to the common surroundings, thereby grounding them in visual experience. The model must then recognize what information Watson is asking for and provide that information as a predicate-based "meaning" that is grounded in the environment.

[1] Corresponding Author: dmonner@cs.umd.edu

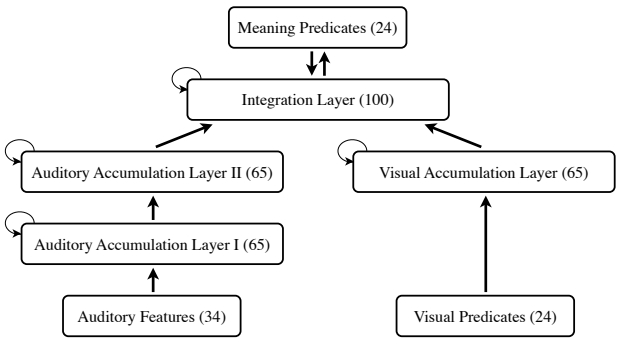

Figure 1. The network architecture of the model. Layers of memory cells are shown with a self-recurrent arc denoting the self-connections of the individual units therein.

2. Methods

The model is situated in a shared micro-world environment with two other speakers, Watson and Sherlock. A training trial begins with Watson asking a question, the answer to which is directly observable in the environment. Sherlock observes his surroundings and gives the answer, which our model observes and attempts to mimic. Sherlock provides each answer as a sequence of grounded predicates which correspond to the meaning of a complete-sentence answer to Watson's question. The model must learn this behavior by imitating Sherlock. We realize that, in real human language learning, Sherlock's raw meaning would not generally be available for the model to imitate (though there is some evidence that the listener may often be able to infer the meaning [2]). However, we still find it useful to examine this limited model, as its predicate-based outputs provide direct evidence that neural models can learn to produce a fully grounded meaning representation of an answer to a question.

The model (Figure 1) is built using the Long Short Term Memory (LSTM) [3,4] architecture, in which units in hidden layers are replaced by self-recurrent *memory cells* protected by multiplicative gates. This type of network can be trained by gradient descent methods similar to those utilized by other recurrent neural networks, and we adopt such a method in this work. Called the LSTM-g training algorithm [5], this variant of gradient descent focuses on maintaining information locality as a means of approaching biological plausibility. Like many gradient-descent based algorithms, LSTM-g utilizes back-propagating errors; however, in light of work showing how such error propagation can be equivalent to Hebbian training methods [6], we argue that back-propagating errors are merely a computationally expedient implementation detail rather than a feature which detracts from the biological realism of the LSTM-g training algorithm.

The network has separate input pathways for accumulating auditory and visual input into gestalt-type representations [7]. The two pathways differ in length because pilot experiments showed that the visual and auditory gestalt representations are best formed with one and two layers of intermediary processing, respectively. After both input gestalts have been formed, they are integrated by another layer of memory cells, which is then used to sequentially generate predicates that specify the grounded meaning output which answers the input question. The network's previous output is fed back to the integration layer to provide a more specific context for sequential processing. The

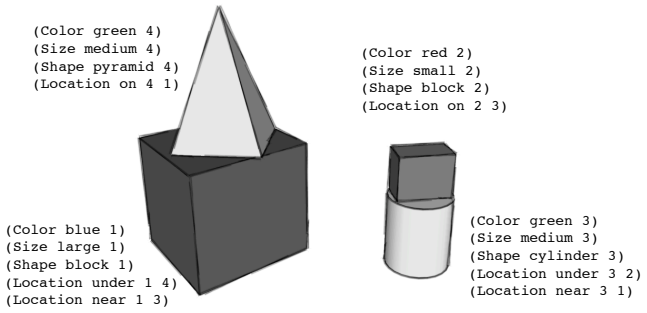

Figure 2. An example of a micro-world environment with four objects, along with the complete set of predicates that describe the environment to the model.

following sections will provide more detail on the forms taken by the auditory and visual inputs and the predicate outputs.

2.1. Environment Input

The micro-world environment shared by our three participants consists of a number of objects placed in relation to each other (see Figure 2). Each object has values for the three attributes size, shape, and color, and each attribute has three possible values: small, medium and large for size; block, cylinder, and pyramid for shape; and red, green, and blue for color. Thus there are 27 distinct objects possible. In addition, each object has a number that identifies it uniquely in the environment, which is a useful handle for a specific object and becomes necessary in the event that the participants need to distinguish between two otherwise identical objects.

Each object is presented to the model as three predicates—coded as binary neural features—with each predicate binding an attribute value to an object identifier. For example, a small red block with unique identifier 2 is completedly described by the predicates `(Size small 2)`, `(Color red 2)`, and `(Shape block 2)`. An environment consist of two to four randomly generated objects, and the predicates describing all these objects are presented at the visual input layer of the model as a temporal sequence at the start of a trial. Each predicate is represented as a collection of active units representing the predicate type, attribute value, and identifying number(s) involved.

Additional predicates are used to describe spatial relations between the objects. One object may be near, on, or underneath another. For example, if our small red block from above (with identifier 2) is on top of a medium-sized cylinder (identifier 3), that fact would be presented to the model as the predicate `(Location on 2 3)`. In our micro-world the on and under relations are complementary (meaning that `(Location on 2 3)` implies `(Location under 3 2)`) and the near relation is symmetric (such that `(Location near 1 3)` implies `(Location near 3 1)`). The location predicates are presented along with the attribute predicates and at the same visual input layer.

Though this space of possible environments may seem small at first, the number of unique environmental configurations is quite large. Using only two, three, or four objects at a time gives us approximately 2.48×10^{10} distinct possible micro-world configurations.

In terms of neural representation, predicate types, attribute values, and identifying numbers each correspond to single units—a drastic simplification of real visual input that

Question	→	where **Is Object** | **What Is Property**
Answer	→	**Object Is Property**
Object	→	the **[Size] [Color] Shape**
What	→	what color **[Size] Shape** | what size **[Color] Shape** | what **[Size] [Color]** thing
Property	→	**Location** | **Color** | **Size**
Location	→	on **Object** | under **Object** | near **Object**
Is	→	is | are *(must agree in number with subject)*
Size	→	small | medium | large
Color	→	red | blue | green
Shape	→	things | pyramid[s] | block[s] | cylinder[s]

Figure 3. The mildly context-sensitive grammar used in the question-answering task. Terminals begin with a lowercase letter while non-terminals are in boldface and begin with an uppercase letter. The symbol '|' separates alternative derivations, and terms in brackets are optional.

we adopt for reasons of computational tractability and because it is a plausible level of representation that might be developed by later stages of the human visual system.

2.2. Question Input

In order to assist in training our model, Watson produces complete English sentences that ask questions about the current shared environment. There are many possible questions for each environment; for the example environment in Figure 2, Watson might ask "What color block is the green pyramid on?", "What thing is under the small block?", "What color are the medium things?", or "Where is the pyramid?".

To produce a question, Watson examines the environment and derives a question beginning from the **Question** non-terminal in the grammar of Figure 3. The question will either be a "what" question asking about one specific property of an object, or a much more complex "where" question requiring a relational predicate and a complete three-predicate object description as an answer. Any objects that Watson refers to must be present in the environment and, to a sophisticated observer, unambiguously identified by the full context of the question. This is necessary because Sherlock must be able to determine the answer to provide a useful imitation target for the model. For example, Watson could not ask "What color is the block?" because it is not clear which block he is referring to, and thus any answer would be poorly defined. Note that Watson can, however, ask questions about groups of objects that share a property, as "What color are the medium things?"—in this case the medium things are the cylinder and pyramid, which are both green, so the answer is well defined. Similarly, questions such as "What color pyramid is red?" that reveal the property being asked about are disallowed.

The words in Watson's question are phonetically transcribed and appended, resulting in a sequence of phonemes that is presented temporally to the model. Each such phoneme is represented as a bundle of binary acoustic features (from [8]) such as voicing, affrication, and formant frequency categories, with each feature corresponding to an input unit. Word and morpheme boundaries are not marked in the input sequence, leaving the model to discover those on its own.

2.3. Answer Output

After Watson has finished asking a question, the model attempts a response. On training trials, Sherlock then gives the correct response, and the model adjusts its connection weights to better imitate this response in the future. Training is thus self-supervised—the

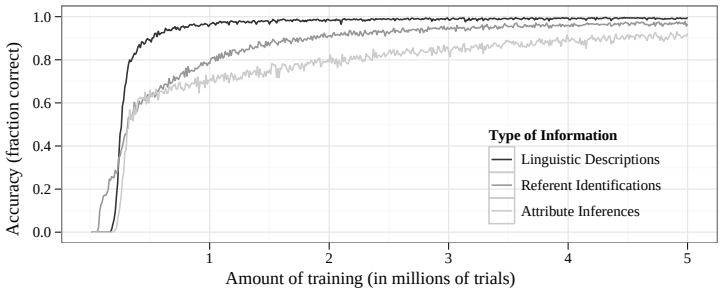

Figure 4. Accuracy versus training time for three separate components of the question-answering task, averaged over all ten training runs.

model need only observe the naturally occurring dialog between Watson and Sherlock and imitate the latter.

To answer Watson's question, Sherlock examines the environment and determines the set of object attributes and relations Watson was referring to. If the question was "What color block is the green pyramid on?" in the context of the environment of Figure 2, Sherlock would produce a sequence of five predicates that correspond roughly to the meaning of the answer sentence "The green pyramid is on the blue block": `(Color green 4)`, `(Shape pyramid 4)`, `(Location on 4 1)`, `(Color blue 1)`, and `(Shape block 1)`.

Note that the answers to the questions differ based on the micro-world environment, such that we never have a one-to-one mapping of questions to answers. Indeed, the answer to Watson's question above would change if the block under the green pyramid were red instead of blue. It would also change (albeit slightly) if the scene were visually identical but the object numbering was altered. Though the above example suggests that the number of possible answer variations is small, some questions admit much greater environment-based variation than others. More open-ended questions such as "where is the block?" could have literally hundreds of possible answers depending on the environment. Thus, neither the model nor Sherlock could answer questions reliably without integrating information from the visual environment.

3. Results

We trained 10 separate networks, each with randomly-chosen weights, on the question-answering task for 5 million randomly generated trials—far less than 0.1% of the input space. All reported results come from evaluations on test trials consisting of a novel micro-world/question pairing. On average, the trained networks were able to produce a perfect temporal sequence of up to 8 grounded predicates to answer the question on 92.5% of these novel trials, strongly suggesting that observe-and-imitate training is sufficient even for the difficult task of grounded question comprehension and answering.

To evaluate the relative difficulty of the components of the task, we can decompose the output predicates into three categories: attribute values corresponding to linguistic descriptions which were present in the question, referent identifications made by grounding the question in the environment, and attributes that are inferred from the environment to answer the question. Comparing the time-course of accuracy over these three categories

in Figure 4, we see that the model learns to understand spoken descriptions fairly quickly, with performance leveling off above 95% after the first 1M trials. At this point, however, performance on referent identifications and question-answers is relatively weak, at 80% and 70% respectively; however, accuracy on these areas continues to slowly improve as training progresses. This seems to us to be compelling evidence that grounded language learning is much more difficult than word recognition, whereas using the environment to infer the answers to questions is harder still.

4. Conclusion

These initial results demonstrate that a purely neural system can learn not only to comprehend language grounded in a shared environment, but also to understand and answer spoken questions. To our knowledge, ours is the first neural network model to successfully learn to map sentence-level language, represented at sub-lexical resolution, onto complex, composable symbolic meanings. We plan, in future work, to investigate the types of representations that the model develops when integrating the auditory and visual streams, and whether a symbol-like representation of the input question can be said to exist independently of the model's derived answer to the question. We are also developing a variant of this model that learns by observing answers from Sherlock as speech streams rather than collections of predicates, with the former being an imitatable signal that is always observable in human language learning scenarios. Such a model would naturally learn speech production capabilities by observation, just as the model presented in this paper suggests that simple observation and imitation might be sufficient tools for learning challenges as complex language acquisition.

Acknowledgments

This work was supported in part by NIH award HD064653.

References

[1] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. Building Watson : An Overview of the DeepQA Project. *AI Magazine*, 31(3):59–79, 2010.

[2] M. Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.

[3] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, Nov. 1997.

[4] F. A. Gers and F. Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, Oct. 2000.

[5] D. Monner and J. A. Reggia. A Generalized LSTM-Like Training Algorithm for Second-Order Recurrent Neural Networks. *Neural Networks (in press)*, 2011.

[6] X. Xie and H. S. Seung. Equivalence of Backpropagation and Contrastive Hebbian Learning in a Layered Network. *Neural Computation*, 15(2):441–54, Feb. 2003.

[7] M. F. St. John and J. L. McClelland. Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46(1-2):217–257, 1990.

[8] S. A. Weems and J. A. Reggia. Simulating Single Word Processing in the Classic Aphasia Syndromes Based on the Wernicke-Lichtheim-Geschwind Theory. *Brain and Language*, 98(3):291–309, 2006.