

The Developmental Approach to Evaluating Artificial Intelligence—A Proposal

Anat Treister-Goren* and Jason Hutchens
Artificial Intelligence NV

ABSTRACT

We propose a developmental evaluation procedure for artificial intelligence¹ that is based on two assumptions: that the Turing Test provides a sufficient subjective measure of artificial intelligence, and that any system capable of passing the Turing Test will necessarily incorporate behavioristic learning techniques.

KEYWORDS: *artificial intelligence, human-computer conversation, Turing Test, child machine, verbal behavior, Markov modeling, information theory*

1. INTRODUCTION

In 1950 Alan Turing considered the question “Can machines think?” Turing’s answer to this question was to define the meaning of the term ‘think’ in terms of a conversational scenario, whereby if an interrogator cannot reliably distinguish between a machine and a human based solely on their conversational ability, then the machine could be said to be thinking [1]. Originally called the imitation game, this procedure is nowadays referred to as the Turing Test.

The field of artificial intelligence (AI) has largely ignored this strict evaluation criterion. Today AI encompasses topics such as intelligent agents, chatterbots, pattern recognition systems, voice recognition systems and expert systems, with applications in medicine, finance, entertainment, business and manufacturing. It could be said that the field is currently in a contentious state. Even though important work has been conducted in terms of the sophistication and expertise of programs, the vision which motivated the birth of AI has not yet been fulfilled: there is neither sufficient cooperation nor agreement amongst its researchers. The unfortunate result of this trend is that true advancement is inhibited. We believe that a new approach is required.

In this paper we shall demonstrate that the Turing Test is a sufficient evaluation criteria for artificial intelligence provided that the expectation level of the interrogator is set appropriately. We propose to achieve this by complementing the Turing Test with objective developmental evaluation. The logical flow of this paper reflects the necessary steps one must take when trying to establish

evaluation standards for artificial intelligence: we begin with a definition of artificial intelligence, we continue with a discussion of the theory and methods which we believe are an essential prerequisite for the emergence of artificial intelligence and we conclude with our proposed evaluation procedure.

2. THE TURING TEST

The Turing Test is an appealing measure of artificial intelligence because, as Turing himself writes, it ...

... has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man.

The Loebner Contest, held annually since 1991, is an instantiation of the Turing Test [2]. The sophistication and performance of computer programs entered into the contest, or lack thereof, bears out our introductory remark that the Turing Test has been largely ignored by the field. In a recent thorough review of conversational systems, Hasida and Den emphasize the absurdity of performance in the Loebner Contest [3]. They assert that since the Turing Test requires that systems “talk like people”, and since no system currently meets this requirement, the *ad-hoc* techniques which the Loebner Contest subsequently encourages make little contribution to the advancement of dialog technology.

Although we agree wholeheartedly that the Loebner Contest has failed to contribute to the advancement of artificial intelligence, we do believe that the Turing Test is an appropriate evaluation criteria, and therefore our approach equates artificial intelligence with conversational skills. We further believe that engaging in domain-unrestricted conversation is the most critical evidence of intelligence.

2.1. Turing’s Child Machine

Turing concluded his classic paper by theorizing on the design of a computer program which would be capable of passing the Turing Test. He correctly anticipated the limitations of simulating adult level conversation, and proposed that ...

... instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child’s? If this were then subjected to an appropriate course of education one would obtain the adult brain.

*All correspondence should be emailed to anat@a-i.com.

¹We use the term *artificial intelligence* to refer to machine intelligence which exhibits human-like conversational capability. To avoid ambiguity, we shall refer to the field of artificial intelligence as AI throughout.

Turing regarded language as an acquired skill, and recognized the importance of avoiding the hard-wiring of the computer program wherever possible. He viewed language learning in a behavioristic light, and believed that the language channel, narrow though it may be, is sufficient to transmit the information which the child machine requires in order to acquire language.

It is indeed unfortunate that this promising line of work was mostly abandoned by the field. Today we find ourselves at a crossroads—a paradigm shift is in the air, and many AI researchers are returning to the behavioristic approach that Turing favoured.

2.2. The Traditional Approach

Contrary to Turing’s prediction that at about the turn of the millennium computer programs will participate in the Turing Test so effectively that an average interrogator will have no more than a seventy percent chance of making the right identification after five minutes of questioning, no true conversational systems have yet been produced, and none has passed an unrestricted Turing Test.

This may be due in part to the fact that Turing’s idea of the child machine has remained unexplored—the traditional approach to conversational system design has been to equate language with knowledge, and to hard-wire rules for the generation of conversations. This approach has failed to produce anything more sophisticated than domain-restricted dialog systems which lack the kind of flexibility, openness and capacity to learn that are the very essence of human intelligence. As far as human-like conversational skills are concerned, no system has surpassed toddler level, if at all.

Since the 1950’s, the field of child language research has undergone a revolution, inspired by Chomsky’s transformational grammar [4] on the one hand and Skinner’s behaviorist theory of language [5] on the other. Computational implementations based on the Chomskian philosophy are the norm, and have yielded disappointing results. It is our thesis that true conversational abilities are more easily obtainable via the currently neglected behavioristic approach.

3. VERBAL BEHAVIOR

Behaviorism focuses on the observable and measurable aspects of behavior. Behaviorists search for observable environmental conditions, known as *stimuli*, that co-occur with and predict the appearance of specific behavior, known as *responses* [6]. This is not to say that behaviorists deny the existence of internal mechanisms; they do recognize that studying the physiological basis is necessary for a better understanding of behavior. What behaviorists object to are internal structures or processes with no specific physical correlate inferred from behavior.

Behaviorists therefore object to the kind of grammatical structures proposed by linguists, claiming that these only complicate explanations of language acquisition [7]. They favour a functional rather than a structural approach, focusing on the function of language, the stimuli that evoke verbal behavior and the consequences of language performance. We believe this to be the right approach for the generation of artificial intelligence.

Skinner argues that psycholinguists should ignore traditional categories of linguistic units, and should instead treat language as they would any other behavior. That is, they should search for the functional units as they naturally occur, and then discover the functional relationship that predict their occurrence.

Behaviorism focuses on reinforced training. Since language is regarded as a skill that is not essentially different from any other behavior, generating and understanding speech must therefore be controlled by stimuli from the environment in the form of reinforcement, imitation and successive approximations to mature performance. Skinner takes the extreme position that the speaker is merely a passive recipient of environmental pressures, having no active role in the process of language behavior or development.

According to behaviorists, changes in behavior are explained through the association of stimuli in the environment with certain responses of the organism. The processes of forming such associations are known as *classical conditioning* and *operant conditioning*.

3.1. Classical Conditioning

Classical conditioning accounts for the associations formed between arbitrary verbal stimuli and internal responses or reflexive behavior. In classical conditioning, for example, the word ‘milk’ is learned when the infant’s mother says “milk” before or after feeding, and this word becomes associated with the primary stimulus (the milk itself) to eventually elicit a response similar to the response to the milk. Once a word or a *conditioned stimulus* elicits a *conditioned response*, it can become an *unconditioned stimulus* for modifying the response to another conditioned stimulus. For example, if the new conditioned stimulus ‘bottle’ frequently occurs with the word ‘milk’, it may come to elicit a response similar to that for the word ‘milk’. Words stimulate each other and classical conditioning accounts for the interrelationship of words and word meanings. Classical conditioning is more often used to account for the receptive side of language acquisition.

3.2. Operant Conditioning

Operant conditioning is used to account for changes in voluntary, non-reflexive behavior that arise due to environmental consequences contingent upon that behavior. All behavioristic accounts of language acquisition assume that

children's productive speech develops through differential reinforcers and punishers supplied by environmental agents in a process known as *shaping*. Children's speech that most closely resembles adult speech is rewarded, whereas productions that are meaningless are either ignored or punished. Behaviorists believe that the course of language development is largely determined by the course of training, not maturation, and that the time it takes children to acquire language is a consequence of the limitations of the training techniques. Operant conditioning is used to account for the productive side of language acquisition.

Imitation is another important factor in language acquisition because it allows the laborious shaping of each and every verbal response to be avoided. The process of imitation itself becomes reinforcing and enables rapid learning of complex behaviors.

Behaviorists do not typically credit the child with intentions or meanings, the knowledge of rules or the ability to abstract important properties from the language of the environment. Rather, certain stimuli evoke and strengthen certain responses in the child. The sequence of language acquisition is determined by the most salient environmental stimuli at any point in time, and by the child's past experience with those stimuli. The learning principle of reinforcement is therefore taken to play a major role in the process of language acquisition, and is the one we believe should be used in creating artificial intelligence.

4. THE DEVELOPMENTAL MODEL

We maintain that a behavioristic developmental approach could yield breakthrough results in the creation of artificial intelligence. Programs can be granted the capacity to imitate, to extract implicit rules and to learn from experience, and can be instilled with a drive to constantly improve their performance. Language acquisition can be achieved through successive approximations and positive and negative feedback from the environment. Once given these capabilities, programs should be able to evolve through critical developmental language acquisition milestones in order to reach adult conversational ability.

Human language acquisition milestones are both quantifiable and descriptive, and any system that aims to be conversational can be evaluated as to its analogical human chronological age. Such systems could therefore be assigned an age or a maturity level beside their binary Turing Test assessment of "intelligent" or "not intelligent".

4.1. Success in Other Fields

Developmental principles have enabled evaluation and treatment programs in fields formerly suffering from a lack of organizational and evaluative principles [8], [9], and have been especially useful in areas which border on the question of intelligence. Normative developmental language data has enabled the establishment of diagnostic scales,

evaluation criteria and treatment programs for developmentally delayed populations. In other areas, such as schizophrenic thought disorder, in which clinicians often found themselves unable to capture the communicative problem of patients in order to assess their intelligence level or cognitive capability, let alone to decipher medication treatment effects, the developmental approach has proven to be a powerful tool [10].

5. LANGUAGE MODELING

We are interested in programming a computer to acquire and use language in a way analogous to the behavioristic theory of child language acquisition. In fact, we believe that fairly general information processing mechanisms may aid the acquisition of language by allowing a simple language model, such as a low-order Markov model, to bootstrap itself with higher-level structure.

5.1. Markov Modeling

Claude Shannon, the father of Information Theory, was generating quasi-English text using Markov models in the late 1940's [11]. Such models are able to predict which words are likely to follow a given finite context of words, and this prediction is based on a statistical analysis of observed text. Using Markov models as part of a computational language acquisition system allows us to minimize the number of assumptions we make about the language itself, and to eradicate language-specific hard-wiring of rules and knowledge.

Some behaviorists explain that language is processed as word-sequences, or response-chains, with the words themselves serving as stimulus for their successors [12]. Information theoretic measures may be applied to Markov models to yield analogous behavior, and more sophisticated techniques can model the case where long-distance dependencies exist between the stimulus and the response.

To date, conversation systems based on this approach have been thin on the ground [13], although the technique has been used extensively in related problems, such as speech recognition, text disambiguation and data compression [14].

5.2. Finding Higher-Level Structure

Information theoretic measures may be applied to the predictions made by a Markov model in order to find sequences of symbols and classes of symbols which constitute higher-level structure. For example, in the complete absence of *a priori* knowledge of the language under investigation, a character-level Markov model inferred from English text can easily segment the text into words, while a word-level Markov model inferred from English text may be used to 'discover' syntactic categories [15].

This structure, once found, can be used to bootstrap the Markov model, allowing it to capture structure at even higher levels. A hierarchy of models is thus formed, each of which views the data at a different level of abstraction. Although each level of the hierarchy is formed in a purely bottom-up fashion from the data supplied to it by the level below, the fact that each model provides a top-down view with respect to the models below it allows a feedback process to be applied, whereby interaction between models at adjacent levels of abstraction serves to correct bad generalisations made in the bootstrapping phase.

It is our belief that combining this approach with positive and negative reinforcement is a sensible way of realizing Turing’s vision of a child machine.

6. EVALUATION PROCEDURE

Our proposal is to measure the performance of conversational systems via both subjective methods and objective developmental metrics.²

6.1. Objective Developmental Metrics

The ability to converse is complex, continuous and incremental in nature, and thus we propose to complement our subjective impression of intelligence with objective incremental metrics. Examples of such metrics, which increase quantitatively with age, are:

Vocabulary size: The number of different words spoken.

Mean length of utterance: The mean number of word morphemes spoken per utterance.

Response types: The ability to provide an appropriate sentence form with relevant content in a given conversational context, and the variety of forms used.

Degree of syntactic complexity: For example, the ability to use embedding to make connections between sentences, and to convey ideas.

The use of pronominal and referential forms: The ability to use pronouns and referents appropriately and meaningfully.

These metrics provide an evaluation of progress in conversational capability, with each capturing a specific aspect. Together they enable an understanding of the nature of the critical abilities that contribute toward our desired goal: achieving a subjective judgement of intelligence.

The challenge in creating maturational criteria is in combining these metrics meaningfully. One might expect discrepancies in the development of the different aspects of conversational performance. For example, some systems may utter long, syntactically complex sentences, typical of a child aged five or above, but may lag in terms of the use of pronouns expected at that age. Weighting the various developmental metrics is far from trivial.

²We use the term *metric* in its non-mathematical sense of relating to measurement.

6.2. The Subjective Component

We do not claim that objective evaluation should take precedence over subjective evaluation, just as we do not judge children on the basis of objective measures alone. Subjective judgement is an important if not determining criterion of overall evaluation. We believe that the subjective evaluation of artificial intelligence is best performed within the framework of the Turing Test.

The judgement of intelligence is in the eye of the beholder. Human perception of intelligence is always influenced by the expectation level of the judge toward the person or entity under scrutiny—obviously, intelligence in monkeys, children or university professors will be judged differently. Using objective metrics to evaluate maturity level will help set up the right expectation level to enable a valid subjective judgement to be made.

Accordingly, we propose that suitable developmental metrics be chosen in order to establish a common denominator among various conversational systems so that the expectation level of these systems will be realistic. Given that subjective impression is at the heart of the perception of intelligence, the constant feedback from the subjective evaluation to the objective one will eventually contribute to an optimal evaluation system for perceiving intelligence.

By using the developmental model, computer programs will be evaluated to have a maturity level in relation to their conversational capability. Programs could be at the level of toddlers, children, adolescents or adults depending on their developmental assessment. This approach enables evaluation not only across programs but also within a given program.

7. CONCLUSION

We submit that a developmental approach is a prerequisite to the emergence of intelligent lingual behavior and to the assessment thereof. This approach will help establish standards that are in line with Turing’s understanding of intelligence, and will enable evaluation across systems.

We predict that the current paradigm shift in understanding the concepts of AI and natural language will result in the development of groundbreaking technologies which will pass the Turing Test within the next ten years.

8. REFERENCES

- [1] A.M. Turing, “Computing machinery and intelligence,” in *Collected works of A.M. Turing: Mechanical Intelligence*, D.C. Ince, Ed., chapter 5, pp. 133–160. Elsevier Science Publishers, 1992.
- [2] Stuart M. Shieber, “Lessons from a restricted Turing test,” Available at the Computation and Language e-print server as `cmp-1g/9404002`, 1994.
- [3] K. Hasida and Y. Den, “A synthetic evaluation of dialogue systems,” in *Machine Conversations*, Yorick Wilks, Ed. Kluwer Academic Publishers, 1999.
- [4] Noam Chomsky, *Syntactic Structures*, Mouton, 1975.

- [5] B.F. Skinner, *Verbal Behavior*, Prentice-Hall, 1957.
- [6] R.E. Owens, *Language Development*, Macmillan Publishing Company, 1992.
- [7] G. Whitehurst and B. Zimmerman, "Structure and function: A comparison of two views of development of language and cognition," in *The Functions of Language and Cognition*, G. Whitehurst and B. Zimmerman, Eds. Academic Press, 1979.
- [8] J.B. Gleason, *The Development of Language*, Charles E. Merrill Publishing Company, 1985.
- [9] A. Goren, G. Tucker, and G.M. Ginsberg, "Language dysfunction in schizophrenia," *European Journal of Disorders of Communication*, vol. 31, no. 2, pp. 467–482, 1996.
- [10] A. Goren, "The language deficit in schizophrenia from a developmental perspective," in *The Israeli Association of Speech and Hearing Clinicians*, 1997.
- [11] Claude E. Shannon and Warren Weaver, *The Mathematical theory of Communication*, University of Illinois Press, 1949.
- [12] O.H. Mowrer, *Learning Theory and Symbolic Processes*, Wiley, 1960.
- [13] Jason L. Hutchens, "Introducing MegaHAL," in *NeMLaP3 / CoNLL98 Workshop on Human-Computer Conversation, ACL*, David M. W. Powers, Ed., January 1998, pp. 271–274.
- [14] Eugene Charniak, *Statistical Language Learning*, MIT Press, 1993.
- [15] Jason L. Hutchens, "Finding structure via compression," in *NeMLaP3 / CoNLL98: New Methods in Language Processing and Computational Language Learning, ACL*, David M. W. Powers, Ed., January 1998, pp. 79–82.