Everyone's a Critic: Memory Models and Uses for an Artificial Turing Judge

W. Joseph MacInnes[1], Blair C. Armstrong[2], Dwayne Pare[3], George S. Cree[3] and Steve Joordens[3]

1. Oculus Info. Inc.
Toronto, Ont. Canada
Joe.macinnes@oculusinfo.com

2. Department of Psychology
Carnegie Mellon University and
The Center for the Neural Basis of Cognition
Pittsburgh, PA, USA

3. Department of Psychology
University of Toronto Scarborough
Toronto, Ont., Canada

Abstract

The Turing test was originally conceived by Alan Turing [20] to determine if a machine had achieved human-level intelligence. Although no longer taken as a comprehensive measure of human intelligence, passing the Turing test remains an interesting challenge as evidenced by the still unclaimed Loebner prize[7], a high profile prize for the first AI to pass a Turing style test. In this paper, we sketch the development of an artificial "Turing judge" capable of critically evaluating the likelihood that a stream of discourse was generated by a human or a computer. The knowledge our judge uses to make the assessment comes from a model of human lexical semantic memory known as latent semantic analysis[9]. We provide empirical evidence that our implemented judge is capable of distinguishing between human and computer generated language from the Loebner Turing test competition with a degree of success similar to human judges.

Keywords

Semantic Memory, General Knowledge, Decision Making, Machine learning, Language, Turing test.

Introduction

Even before the formal birth of Artificial Intelligence (AI), it was believed by some that computers would achieve human-level intelligence within a relatively short time, so it was essential to devise a test to determine exactly when this milestone had been reached. To this end, Alan Turing [20] proposed the Turing test as one means of evaluating the intelligence of an artificial entity. In essence, he proposed that a computer could be deemed intelligent if it could believably mimic human communication. Specifically, he proposed a guessing game, played by a human confederate, an artificial entity, and – central to this paper - a judge. Without knowing their true identities, the judge would converse with both the confederate and the artificial entity. If the judge was unable to systematically identify which of the two was human, the artificial entity would be said to be intelligent.

Although the classic Turing test is no longer seen as an acceptable measure of human intelligence[18][17] , it remains an excellent and incredibly difficult test of language mastery. It can also serve as a valid test of agent believability where the standard may only be to mimic human behaviour [15]. Currently, the annual Loebner competition[7] the most renowned forum for attempts at passing the Turing test, has set a more modest threshold for intelligence than the Turing test: only 30% of the judges need to make incorrect attributions of human intelligence for an attribution of intelligence to be made. Nevertheless, this achievement has yet to be accomplished.

This paper will focus on the oft-forgotten third party of the Turing test: the Turing judge. Since it is the objective of the judge to make the determination of whether the intelligence is human or artificial, the task of implementing an artificial judge is simpler than that of creating a artificial contestant – a test of language recognition and understanding, not generation.

The applications of a language judge are many, both within and outside the context of the Turing test. Within the context of the Turing test, we argue that improved AIs would benefit from a component which evaluates the quality of a generated reply. Our argument to this effect is derived in part from evidence within the cognitive psychology and cognitive science literatures indicating that humans employ some form of critic themselves during sentence comprehension and generation – "a reader tries to digest each piece of text as he encounters it" [22, p. 16]. As one salient example, the manner in which humans process 'garden path' sentences[4] whose latter portions do not conform to the interpretation typically expected by the former portion (e.g., the cotton clothing is made of is grown in the South) suggests that we evaluate likely sentence meaning continuously as we read a sentence.

Outside the context of the Turing test, multiple alternative applications abound: evaluation of the quality of student essays[10][19] identification of human versus computer generated on-line forum posts, e-mails, and other forms of web traffic, and the development of security software designed to segregate typical human computer interactions versus automated intrusion attempts.

We have undertaken a principled approach to the development of the first generation of our Turing judge. Our approach draws its inspiration from the early development of artificial intelligence (e.g., Newell & Simon, 1956), which is currently embodied to some extent within the interdisciplinary realm of cognitive science: we aim to advance AI in part through our understanding of human intelligence. Further discussion of this issue awaits later in the paper, but this fact is worthy of emphasis for two reasons: first, it highlights the benefits of a multidisciplinary approach to tackling general AI issues. Second, we wish to explicitly acknowledge that although the "human computer" has been honed over millions of years of evolution, it is clearly lacking in many regards. Future collaborative efforts integrating more non-human approaches in the development of improved Turing judges would therefore be most welcome.

The Turing Judge

The fundamental goal of the Turing judge is to ascertain whether a sentence or passage of text was generated by a human or not. The passage could be evaluated on multiple dimensions: grammaticality (e.g., he throws the ball vs. he throw the ball), meaningfulness of content (e.g., colorless green ideas sleep furiously [2]), relatedness of content to previously discussed content, and so on. Vast literatures and many complex issues surround each of these topics. In developing our first model, we have focused our efforts on two of these issues: assessing the meaningfulness and relatedness of semantic content. These issues in particular seem to be the most fundamentally challenging and relevant to AIs currently being developed to pass the Turing test, as a common strategy in recent years been to simply select a pre-programmed response to a given question from amongst a database of sentences recorded from humans [23].

For the judge to appropriately evaluate the passage of text, it must be supplied with some knowledge of human discourse. To address this issue, we turned to the literature examining the derivation of lexical semantic knowledge (i.e., the derivation of a word's meaning) from how words co-occur within large samples of natural language (corpora of written text). Numerous computational models have been developed aimed at extracting different components of structure from within text, and these models have shown considerable success at accounting for a wide variety of comprehension phenomena. Examples include: assessing the correctness of word order in a section of text [24] and comprehending metaphors [25] among others.

When selecting a particular word co-occurrence model to employ in our judge, two main forces came into play. The first was a model's performance on conversational tasks similar to those a Turing judge might encounter, and the second was the degree to which the model tends to perform well across a wide variety of tasks. Space constraints prevent a detailed discussion of these issues here, but they are expounded in [3]. It suffices to say that consideration of these issues led us to select the Latent Semantic Analysis (LSA [8]) model for use in our Turing judge. It chronologically predates most other models and has been tested in the most diverse set of tasks. It has performed well in most tasks and has been adopted as the de facto benchmark model when comparing the performance of newer models. LSA also has the tangential benefit of being debatably the most well known and easy-to-implement of these models, which should facilitate both the comprehension of the present work, and the execution of future investigations.

Overview of LSA

LSA [8] is a corpus-based statistical method for generating representations that capture aspects of word meaning based on the contexts in which words co-occur. In LSA, the text corpus is first converted into a word x passage matrix, where the passages can be any unit of text (e.g., sentence, paragraph, essay). The elements of the matrix are the frequencies of each target word in each passage (see Figure 1). The element values are typically re-weighted, following a specific mathematical transformation (e.g., log transform) to compensate for disproportionate contributions from high-frequency words. The entire matrix is then submitted to singular value decomposition (SVD), the purpose of which is to abstract a lower dimensional (e.g., 300 dimensions) meaning space in which each word is represented as a vector in this compressed space. In addition to computational efficiency, this smaller matrix tends to better emphasize the similarities amongst words**.** Following the generation of this compressed matrix, representations of existing or new passages can be generated as the average vectors of the words the passage contains**.**

Methods

Our implemented Turing judge used an LSA memory model to assess the meaningfulness and relatedness of discourse. The discourse used at test was from previous attempts at the Loebner competition, so as to determine whether the model can accurately distinguish human generated and computer generated responses. Our hypothesis was that human judges use (at least in part) a measure of the semantic relatedness of an answer to a question to spot the computers, so a model which has these strengths should perform fairly well.

LSA Training

The first step in building an LSA model is to select the text database from which the word matrix will be built. Selecting appropriate training data presents a challenge as the questions posed in the Turing test are completely open ended and can be about any topic. As with many machine learning algorithms, the quality of the semantic representations generated by the LSA model often comes down to a question of quantity versus quality of training data. Ultimately, Wikipedia was chosen due to the online encyclopaedia's aim of providing a comprehensive knowledgebase of virtually all aspects of human knowledge, and for its similarity to the training corpora typically used to train word co-occurrence models. It was hoped that the large volume of information in the Wikipedia corpus would compensate for the lack of question and answer style dialogue (as is present in the Live Journal website), although we intend to revisit the trade-offs associated with each of these alternatives in the future.

The entire June 2005 version of Wikipedia was used as a training set for our instantiation of LSA. This corpus contained approximately 120 million words stored in approximately 800 000 unique articles. Each article was pre-processed to remove all of its html and Wikipedia mark-up, so as to generate a "what you see is what you get" version of the database from which LSA could learn. These articles were further stripped of all of their non-alphanumeric characters, all words were converted to lowercase, and function words such as 'the' and 'that' were trimmed because their high frequency ("the" occurs about once every ten words in the average English sentence) and low meaning content tend to detract from LSA's performance.

To illustrate the judging process, consider how the judge would evaluate the similarity of the question "The humans built what?" relative to the responses "The humans built the Cylons" and "They built the Galactica and the vipers", in the case where the judge had access to the simplified LSA memory model outlined in Table 1 (this example matrix forgoes the SVD compression of the matrix for ease of interpretation, but this transformation would be applied in the full version of the model). First, it would combine the vector representations of each of words (i.e., the rows from the matrix) in each sentence to form a vector representing the combined meaning of each of these words. Ignoring words not present in LSA's memory, the question vector v_q would be equal to (v_{human} + v_{build}), and the answer vectors v_{a1} and v_{a2} would be equal to (v_{human} + v_{build} + v_{cylon}) and (v_{built} + $v_{galactica}$ + v_{viper}) respectively. Note that all of the component vectors which make up v_q and v_{a1} point in roughly the same direction in LSA's memory space (the human-building-cylon region), whereas the component vectors in v_{a2} tend to point to a different region of space than the question vector (the survivors-with-vipers-on-galactica region). Consequently, v_q and v_{a1} would have a higher cosine value, and v_{a1} would be considered the better or more "human" answer.

A1. Humans built the Cylons to make their lives easier.
A2.The Cylons did not like doing work for the humans.
A3.In a surprise attack, the Cylons destroyed the humans that built them.
A4.The Cylons were built by humans to do arduous work.

B1. Some survivors escaped and fled on the Galactica.
B2. The Galactica protected the survivors using its Viper attack ships.
B3. The Cylons were no match for a Viper flown by one of the survivors.
B4. A Viper flown by one of the survivors found Earth and led the Galactica there.

	A1	A2	A3	A4	B1	B2	B3	B4
built	1		1	1				
cylons	1	1	1	1			1	
humans	1	1	1					
a			1				1	1
galactica					1	1		1
survivors					1	1	1	1
viper						1	1	1

Table 1. Simplified LSA representation for the eight sentences listed above. Each column represents a sentence, and each row represents how frequently each word occurred in that sentence. In this example, words which did not occur at least three times across all sentences and the entry for the function word 'the' have been removed from the table. This matrix has also not been subject to singular value decomposition so as to render it more conceptually straightforward to interpret, although this procedure would be applied in the full model. Note that although LSA has no other knowledge about the world, it nevertheless captures the fact that ('humans', 'built', and 'cylons'), and ('galactica', 'survivors', and'viper') form clusters of meaningfully related knowledge, and that these clusters are largely separated from one another.

LSA Turing Judge Performance

We aimed to evaluate the similarity between the Judge's questions relative to both the AI and the human answers in

previous Loebner prize conversations. To do so, we first compiled each conversation into question and answer pairs: the judge's question followed by the answer of the conversational agent**.** Our artificial judge then queried the LSA model and had it return the high-dimensional memory vector corresponding to each word in each of the questions and answers. The vectors for the words comprising the questions and answers were then separately conflated to derive a separate representation of the question and the answer in the high-dimensional memory space provided by LSA. The cosine similarity of these vectors was then calculated and used as the metric for the "humanness" of the human or AI agent in question.

Our hypothesis was that a human agent, by virtue of their better overall conversation ability, would have higher semantic similarity with the human judge's question than any of the artificial agents. Furthermore, we hypothesized that our LSA judge would employ a metric similar to human judges in assessing whether the agent was a human or not. Consequently, the rank ordering of the different agents provided by our artificial judge should correspond with those of the human judges. To assess the validity of these hypotheses, we used our judge to evaluate the humanness of the artificial and human agent discourse with the human Turing judge from the 2005 Loebner competition. There were approximately 35 question-answer pairs tested for each of the AIs, and 135 question-answer pairs tested for the humans; humans having more data points because they participated along with each AI in each run of the Turing test.

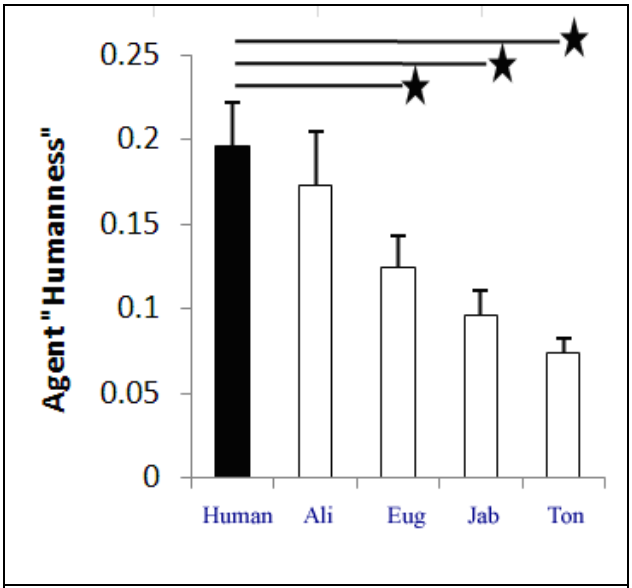

Figure 1. The artificial Turing judge's "humanness" rating for both the human agent (black bar) and the artificial agents (white bars; Ali → Alice, Eug → Eugene, Jab → Jabberwacky, Ton → Toni). "Humanness" was operationalized as the cosine of the conflated LSA vector similarity for all of the words in the question relative to all of the words in the answer. Error bars are the standard error of the mean. Statistically significant differences between the human agent and the artificial agents are denoted with stars (see text for details). With the exception of ALICE, humans scored significantly higher than the artificial agents.

Results

Based on the process outlined above, our artificial Turing judge generated a 'humanness' rating for the human and artificial intelligences and these are reported in Figure 1. As predicted, humans were rated as most "human" by our judge, with each of the artificial agents showing lower performance relative to actual humans. We subjected the humanness ratings to one-way analysis of variance (ANOVA), and pair-wise t-tests[1] of the human agent against all of the artificial agents. A significance threshold of p = .05 was used in all analyses. These analyses indicated significant overall differences between the conditions ($F(4,262)$ = 2.7), and the pair-wise t-tests indicated that the human agent was rated as significantly more human than all of the AIs except for ALICE (Human vs. ALICE $t(88.6)$ = 1.6; Human vs. EUGENE $t(142.8)$ =

Artificial Judge	**Human Judge 1**	**Human Judge 2**	**Human Judge 3**	**Human Judge 4**
Hum (.20)	Hum (77)	Hum (75)	Hum (79)	*Hum (88)*
Ali (.17)	Jab (40)	Eug (40)	Eug (45)	*Eug (30)*
Eug (.12)	Eug (35)	Ton (30)	Jab (20)	*Ali (10)*
Jab (.10)	Ton (10)	Ali (20)	Ton (15)	*Jab (8)*
Ton (.07)	Ali (9)	Eug (10)	Ali (5)	*Ton (2)*

Table 2. Rank orderings and performance scores of the different artificial agents as determined by our artificial Turing judge and the four human judges who evaluated the agents during the Loebner competition. Note both the similarity between our artificial judge's ratings and those of the fourth human judge (both in italics), and the substantial variability in the rank orderings of the different agents by the different human judges (Hum → Human, Eug → Eugene, Jab → Jabberwacky, Ton → Toni).

[1] Given that large differences in the variability of the humanness ratings for the different agents, equal variance was not assumed when running the t-tests; hence, a separate estimate of each condition's variance and adjusted degrees of freedom was used to compensate for violating the t-test's homogeneity of variance assumption.

2.7; Human vs. JabberWacky $t(157.6) = 3.4$; Human vs. Tony $t(157.1) = 4.5$).

To further assess the performance of our artificial Turing judge, we investigated how well our judge's approximation compared to the humanness metric used by actual human judges. To do so, we compared the ordinal rank orderings of the artificial agents in terms of humanness as determined by our artificial Turing judge against the ordinal rank orderings generated by the human judges during the Loebner competition. These data are presented in Table 2. First, on a qualitative level our artificial judge's rank orderings (first column) are quite similar to those of the fourth human judge (the two top rated agents being interchanged across the two judges). Second, there is considerable variability in the rank ordering of the different agents across the different judges.

More formally, we examined the correlations amongst the raw scores provided by each of the judges so as to determine the average consistency across the different judges. These analyses showed that there are significant differences in terms of the average correlation amongst the human judges (mean correlation = .83, SE = .045) and amongst each of the human judges and the artificial judge (mean correlation = .59, SE = .06), $t(8) = 3.34$. Thus, in the details there is clearly room for improvement in our artificial judge, primarily in terms of rating humans as being vastly more "human" than their AI counterparts. Nevertheless, our mean human judge versus artificial judge correlation of .59 is quite substantial (reaching a maximum of .76 amongst the artificial judge and the fourth human judge), and provides at least modest support for the conceptual validity of our approach.

Discussion

This work demonstrates that an artificial Turing judge with access to lexical semantic representations such as those derived by LSA is capable of distinguishing human and computer generated conversation agents with a high degree of accuracy. This test bodes well for semantic detectors as a key component of a more comprehensive artificial Turing judge capable of making more robust and sensitive discriminations. Moreover, the failing of most artificial agents to achieve "human" level semantic similarity amongst the question and responses indicates that enhancing the meaningfulness and relatedness of the answers artificial agents provide to questions they are posed warrants substantial attention by AI researchers interested in the Turing test and related issues.

Despite our model's success, we note several means in which it could be enhanced. For instance, it has yet to be determined whether LSA represents the best knowledge base for the Turing judge to probe when evaluating the humanness of a sentence, nor whether the usage of the cosine is the best metric for assessing the similarity of the content of two passages of text (see [26] for discussion). Furthermore, there are clearly many other dimensions of humanness of a text passage which the current judge ignores (e.g., grammaticality). Framed in a broader context, we view the present work as demonstrating the validity and potential of an artificial Turing judge and the importance semantic knowledge plays in assessing 'humanness'. Nevertheless, there is much which remains unexplored in developing this oft-neglected subcomponent of the Turing test.

Next steps will include a comparison of the current critic trained on Wikipedia with a second critic trained on Live Journal conversations to determine if the conversational style corpus helps in a conversational critic. Live journal offers interesting potential for Turing judges. Since the data is organized by the on-line persona which authored the text, we have an excellent opportunity to train algorithms which also exhibit certain personalities. Each persona contains an accessible description of its author's personality along with a keyword list of user interests. Using these lists, it is quite feasible to train an algorithm with personas interested in a particular topic. For example, we could train algorithms from personas interested in anthropology, computers, or swimming, and in theory, the algorithms may learn more from these areas than others.

Conclusion

Ultimately, for any system to perform the Turing test at a high level it will have to combine information from a variety of sources, and choose among a number of potential responses supported by these sources. Some form of internal judge or critic could be critical in this regard. The current research is the first stage in an interdisciplinary project designed to model human cognition. As we improve our techniques to more human-level computer interaction, we will also need to consider our methods for assessing those techniques. Self-evaluation processes are likely critical to efficient human performance in a wide range of problem solving contexts. The Turing test provides a clearly defined context in which to create and test such self-evaluation processes, and modelling the judge seems to us to be a very reasonable starting point in this regard, and a useful task in its own right.

Acknowledgements

This work was supported by a University of Toronto Scarborough Academic Initiative Fund Grant to WJM, National Sciences and Engineering Research Council (NSERC) Alexander Graham Bell Canada Graduate Scholarship to BCA, and an NSERC Discovery Grant to GSC and WJM.

References

[1] Burgess, C., & Lund, K. (1997). Parsing constraints and high dimensional semantic space. *Language & Cognitive Processes*, 12, 177-210.

[2] Chomsky, N. (1957). *Syntactic Structures.* Mouton: The Hague.

[3] Cree, G. S., & Armstrong, B. (in press). Computational models of semantic memory. In M. Spivey, K. McRae, & M. Joanisse, *The Cambridge Handbook of Psycholinguistics*.

[4] Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 25, 725–745

[5] Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25, 2&3, 285-307.

[6] Live Journal website. Http://www.livejournal.com

[7] Home page for the Loebner prize, a current implementation of the Turing test. Http://www.loebner.net/Prizef/loebner-prize.html

[8] Landauer, T. K., Foltz, P. W., & Laham, D. (1998).Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

[9] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

[10] Landauer T.K., Laham D. & Foltz P. (2003) Automatic essay

[11] assessment. Assessment in Education, Principles, Policy &

[12] Practice 10, 295–308.

[13] Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203-208.

[14] Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high dimensional semantic space. *Proceedings of the Cognitive Science Society*. 660-665. Hillsdale, NJ: Erlbaum.

[15] MacInnes, W.J. (2004) Believability in Multi-Agent Computer Games: Revisiting the Turing Test. *Proceedings of CHI*, 1537.

[16] Newell, A., & Simon, H. A. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*.

[17] Nilsson, N. J. (2005). Human-level artificial intelligence? Be serious!. *AI Magazine, 26*(4), 68-75.

[18] Oppy, G., & Dowe, D. (2003). The Turing Test. In E. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Available online at http://plato.stanford.edu/entries/turing-test/.

[19] Pare, D. E., & Joordens, S. (2008). Peering into large lectures: Examining peer and expert mark agreement using peerScholar, an online peer assessment tool. *The Journal of Computer Assisted Learning*, 24, 526-540.

[20] Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 236, P433.

[21] Walter Kintsch (2001). Predication, *Cognitive Science* 25, 173–202

[22] Just, M. A., & Carpenter, P. A. (1987). The Psychology of Reading and Language Comprehension. Allyn and Bacon, Inc: Newton, MA.

[23] http://loebner.net/SDJ_Interview.html

[24] Landauer. T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent Semantic Analysis and humans. In M. G. Shafto & P. Langley, (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, pp. 214-417. Mahwah, NJ: Erlbaum.

[25] Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, 7, 257-266.

[26] Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2007). An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Science*, Submitted.