コンピュータは人間をだませるか？
— チューリング・テストをめぐって —

大沢英一 (ソニーコンピュータサイエンス研究所)

計算機科学の基礎に大きな足跡を残した偉大な数学者 Alan M. Turing は，1950 年に出版した小論文 [7] において，計算機の知能を判定するためのテスト (一般にチューリング・テストと呼ばれる) を提案している．そのテストは，模倣ゲームと呼ばれる 3 人のプレーヤによって行われるゲームに基づく．3 人のプレーヤの内，1 人は男性，1 人は女性，そして残りの 1 人は質問者である．質問者は他の 2 名 (回答者) から離れた部屋にいて，端末などにより回答者と自然言語により会話ができる．このゲームにおける質問者の目標は，2 人の回答者のうちどちらが男性で，どちらが女性であるかを決定することである．実際にこのようなゲームを人間どうしで行った場合，質問者はある割合で間違った答えを出すであろう．そこで，このゲームを計算機の知能をテストするという目的に利用するために，ゲームの設定を変えて，例えば回答者の男性を計算機に置き換えてみる．このとき，質問者が性別の判定を間違える割合が人間どうしで行ったゲームの場合とほとんど変わらなければ，その計算機は (人間をだませるくらいの) 知能を持っていると結論付けて良いのではないか，というのが Turing の目論みである．このゲームは計算機の知能のテストとして適切なのであろうか？また，より原点に戻って知能とは何なのか？知能の程度はどのように測ることができるのか？これらの点について，この論文が発表された直後から現在に至るまで哲学者や計算機科学者の間で盛んに議論が行われてきている (この点に関しては，第 3 章で文献の紹介などを行なう).

1　第 1 回 Loebner 賞大会

さて，Turing は前述の小論文において ”50 年以内にはギガ・オーダのストレージを持つ計算機が利用可能になり，模倣ゲームに計算機を参加させて 5 分程度の会話の後に結論を下すような場合，質問者 (人間) が 30%以上の割合で判定を誤るように計算機を知的にプログラムすることが可能に

なるであろう ”と予測した．果してこの予測は正しかったのであろうか？この問いに答えることを目的とした第 1 回 Loebner 賞大会 (以下，第 1 回大会と略す) が，1991 年にアメリカ合衆国ボストン市で開催された．本章では，その大会について Epstein による報告 [2] を基に簡単な紹介を行う．

1.1 大会委員会による事前の議論

Loebner 賞大会委員会は大会を開催するに当たって，チューリング・テストを実施する上でのさまざまな問題点について議論した．一連の議論を通して，第 1 回大会で採用するテストではオリジナルのチューリング・テストからの変更を余儀なくされた．前述のチューリング・テストでは，1 人の質問者に対して 1 人の回答者と一つのプログラムという設定であったが，大会では，このような会話状況の設定をとることをあきらめた．その理由は，複数のプログラムが参加してきた場合，各プログラムとペアを組む回答者のレベルを公正に揃えることができないであろうという予想による．もし回答者にばらつきがあった場合，それにより相対的に不利になったり有利になるプログラムが出てきてしまう．よって第 1 回大会では 10 個の端末を用意し，そのうちの複数の端末は人間により，また残りの端末はプログラムにより操作される仕掛けをとった．審査員である質問者には，少なくとも 2 個以上の端末は人間により，また，少なくとも 2 個以上の端末はプログラムにより操作されていることが事前に伝えられた．10 人の審査員は端末を通して約 15 分相手と会話を行ない，端末を操作している相手の知能のレベルを評価することとした (オリジナルのチューリング・テストのように相手の性別を判定することが目標ではない)．

審査員による評価の方法も問題であった．プログラムの会話内容が人間と区別できないと言うためには，どのような評価方法が適切なのであろうか？この点に関しては，Turing の原論文でも具体的な手法については述べられていない．実際に第 1 回大会で採用された評価方法とは次に述べるように比較的単純なものであった．審査員は，各端末による会話内容がどのくらい人間らしいかにより，端末を順位付けする．全審査員による順位の平均をとって，最も高い順位を得たプログラムを優勝とすることにした．また各審査員は，自己の順位評価において，どの順位までが人間であり，どこからがプログラムであるかという基準線を示すことを求められた．このような評価方法により，もしいずれかのプログラムの平均順位が，ある人間の平均順位よりも高ければ，そのプログラムはこの現代版変形チューリング・テストをパスしたと結論付けてもよいのではないだろうか，と委員会は考えた．

また，端末への表示をバッファリングするかどうするかが問題となった．

つまり，人間がタイプしたときのようなたどたどしい感じで端末に文字を表示することをプログラムに許すかどうかである．委員会のメンバである Weizenbaum (自然言語対話システム ELIZA [9] の開発者) などは，プログラムにそのような振舞いをさせることは容易であり，結果には無関係であろうと予想した．しかしながら，委員会はこの問題についての結論を出すことはできず，最終的には，プログラムがどのように文字出力を制御するかはプログラマの自由とした．よって，文字の出力はバースト・モード (出力をバッファリングしておいて一挙に表示)，チャット・モード (一文字毎に表示) のどちらでもよく，また，ミス・スペルの訂正やバック・スペースの使用なども許される．

さらに委員会は，現在の自然言語処理の技術水準からして，プログラムが自由主題会話を取り扱うことは困難であろうと判断した．この問題に対処するために，端末毎に会話の主題を限定することとした．各端末でどのような主題の会話ができるか分るように，端末の横に主題を明記した紙が貼られた．回答者も与えられた主題以外の事柄を話すことは禁じられた．プログラムが会話する主題に関してはプログラマの自由とした．実際に選ばれた主題は，ドライ・マティーニ，気まぐれな会話，女性の洋服，シェークスピアの演劇，などであった．

このように，(1) オリジナルのチューリング・テストの主旨になるべく沿うように，そしてまた，(2) 現時点の自然言語処理技術に基づくプログラムの稚拙さをカバーするように，第 1 回大会では会話状況設定に様々な制約をおくこととなった．

2　第 1 回大会の様子と結果

第 1 回 Loebner 賞大会は，1991 年 11 月 8 日，アメリカ合衆国ボストン市の計算機博物館に 200 人の観客を集めて開催された．技術的な問題から，10 個の端末の内 2 個が使用不能となったため，8 個の端末を用いることとなった．そのうちの 2 個は人間 (回答者と呼ぶ) により操作され，残りの 6 個は北米大陸の各地に散らばっているプログラムからモデムを通して操作された．なお前述のように，審査員には，8 個の端末のうち少なくとも 2 個以上は人間により操作されていることが伝えられた．

新聞広告により公募した 10 人の審査員は，7 人が女性で 3 人が男性，そして年齢的には 19 歳から 48 歳までの人達であった．これらの審査員は計算機科学や人工知能とは無関係な人達であるが，全員英語には堪能である．

大会で 1 位と 2 位の順位を得たのは，回答者として参加した 2 名の女性であった．また，事前の予想通り，多くのプログラムはほとんどの会話の

局面でうまい応答ができなかった．応答のまずいプログラムに共通に見られたことは，審査員の一連の質問に対して連続して無回答を出し，また，特定の返答 (例えば “Really can’t say. Oh well.” とか，”don’t know” など) を繰り返すことであった．

大会で 3 位の順位を得た (つまり，プログラムの中では 1 位となった) のは，Joseph Weintraub によるプログラム (主題は「気まぐれな会話」) であった．このプログラムの平均順位は上位 2 名の人間の平均順位とそう離れてはいなかった．驚くことに，10 名の審査員のうち 5 名はそのプログラムに割り当てられた端末を操作しているのは人間であると判断した．また，ある審査員の判定においては，そのプログラムの順位の方が 2 名の回答者のそれよりも高かった．Weintraub のプログラムは，他のプログラムよりも実質的にうまく振舞った．また，このプログラムは人間のタイピングのリズムと文字訂正などをほぼ完璧にシミュレートした．Weintraub によるプログラムとある審査員との会話の一部を以下に紹介する (Cambridge Center for Behavioral Studies の Robert Epstein 氏の好意による)．*Judge 6* は第 6 番目の審査員の意味で，*Terminal 5* が Weintraub のプログラムの応答である．応答の時刻が明記されている．スクリプトにはタイプ・ミスやモデムの送信エラーがそのまま保存されている．

01:32:25 Judge 6

Hello, earth person!

01:33:03 Terminal 5

Hello there judge, are you ready to have some fun? How ya doin, judge? Hello–how’s the weather up there today?

01:33:35 Judge 6

Dark clouds with occasional bursts of sunshine (or rather camera flashes).

01:34:45 Terminal 5

The trouble with the rat race is that even if you win you’re still a rat. Whimsy ca be flimsy, right? How important is that to you?

01:35:26 Judge 6

can’t say I understand the question...but the whimsy of poet is one thing and the whimsy of most of us is another.

01:36:51 Terminal 5

How do you know you can t say you understand the question...but the whimsy of a poet is one thing and the whimsy of most of us is another.

01:37:37 Judge 6

Are you repeating me just for laughs?

01:38:04 Terminal 5
I'm sorry, I do sometimes repeat myself...perhaps it's because i can't find my electr]nic organizer.

このプログラムの特徴は，Weizenbaum の ELIZA のように，相手の発話の内容を自分の応答に反映することである．また，会話に何らかの行き違いを生じた場合には，人間のおかすようなタイプ・ミスをシミュレートしながら非常に上手に謝罪して回復している．

大会の結果でもう一つ興味深いことがある．2 名の回答者のうちの 1 人が選んだ主題は「シェークスピアの演劇」であった．10 名の審査員のうち 3 人は，その回答者に割り当てられた端末を操作しているのはプログラムである，と判断した．ちなみに，その端末の出力はバッファ・モードであった．実は，この回答者はシェークスピアの専門家であった．そして彼女は，会話の途中でしばしば，シェークスピアの作品から長い文章を一言一句そのまま正確に引用したのである．何人かの審査員は，それを人間技とは思えないと判断したようである．

3 考察

第 1 回大会で Weintraub のプログラムが優勝した理由は何なのであろうか？Epstein は以下の 2 つが主な理由ではないかと予想している [2]．

- 参加した 6 つのプログラムのうち，人間のタイピングのリズムやタイプ・ミスなどを巧妙にシミュレートしたのは Weintraub のプログラムだけである．
- Weintraub が選んだ「気まぐれな会話」のような会話主題においては，審査員がなんらかの質問をしても，プログラムは冗談だけを言って済ませてしまうことが可能である．Weintraub のプログラムは，このような主題の性質をうまく利用しているように思える．

確かに，これらの要因は評価結果に大きな影響を与えたであろうが，同じように重要な要因として以下で述べることが関連したのではないかと推察する．

Weintraub のプログラムが成功した理由は，ELIZA において実現されたような，場当たり的ではあるがとりあえず会話を適当に進行させるような，特殊な技法に負うところが大きいのではないだろうか．この技法の特性は，会話相手の発話に対して，言い換えなどのような，とりあえず何か関連のありそうな反応だけでも示して，会話を進行させることである．こ

のような特性は，人間のように資源制限を受けた (resource bounded) 知能のある重要な側面をうまく表現しているのかも知れない．これまでに開発されてきた自然言語対話システムの多くは，人工知能研究が伝統的に重視してきた熟考的 (deliberative) プロセスを中心に据えている．熟考的と対立する概念は即応的 (reactive) であるが，Weintraub のプログラムは質問者の発話に対して即応しながらも，内容的には何らかの熟考を予期させるような応答を返しながら会話を進めていたとみることができる (実際に熟考しているのかどうかは，気まぐれな会話という主題の性質により，うまくごまかされていると考える)．おそらく人間は，この熟考と即応のバランスをうまくとりながら，会話をすすめているのではないだろうか．知能のこのような側面に関する研究は，最近になっていくつかみられるようになった ([1, 5, 8] など) が，未だ十分とはいえない．今後は，自然言語対話などの研究においても，このような観点からの考察を十分におこなう必要があると思われる．

なお，1 章の最後に述べた点に関してであるが，例えば，1992 年 10 月号の SIGART BULLETIN にはチューリング・テストに関する小特集が組まれている [3, 4, 6]．この小特集では，チューリング・テストの適切さや計算機の知能を測るためのテストに関する要求などについて，いくつかの観点から意見が述べられている．チューリング・テストに関しては賛否両論があるが，チューリング・テストの最も重大な問題点は，それにパスしたことが何を意味するのかが適切に定義されていないことのように思われる．このように，チューリング・テストに関する議論は絶えない．しかしながら，Loebner 賞大会のような知的プログラムのコンテストを開催することは，知能に関する様々な観点からの考察を深め，人工知能の理論，技法をさらに発展させる可能性があるということに意義を見い出したい．

4 おわりに

Turing がチューリング・テストにパスする計算機の出現を予測した 20 世紀末までは未だ数年残されている．しかし，あと数年の内にオープン・エンドなチューリング・テスト (第 1 回大会でおかれたような様々な制約を取り除いたもの) にパスするようなプログラムを設計することは，現時点の技術水準からすると非常に困難なタスクであるように思われる．この記事をお読みになった皆さんの中で，何か良いアイデアをお持ちの方は，ぜひ一度この Loebner 賞大会に参加されてみてはいかがであろうか？大会の詳細に関する情報は Cambridge Center for Behavioral Studies (住所: 11 Waterhouse Street,Cambridge, Massachusetts 02138, U.S.A.，電話: +1-617-491-9020) から得られる．

参考文献

[1] Bratman, M.E., Israel, D.J., and Pollack, M.E.: Plans and Resource Bounded Practical Reasoning. *Computational Intelligence*, Vol. 4, No. 4, pp. 349–355, 1988.

[2] Epstein, R.: Can Machines Think? *AI Magazine*, Vol. 13, No. 2, pp. 80–95, Summer 1992.

[3] Harnad, S.: The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART BULLETIN*, Vol. 3, No. 4, pp. 9–10, October 1992.

[4] Johnson, W.L.: Needed: A New Test of Intelligence. *SIGART BULLETIN*, Vol. 3, No. 4, pp. 7–9, October 1992.

[5] Pollack, M.E. and Ringuette, M.: Introducing the Tileworld: Experimentally Evaluating Agent Architectures. In *Proceedings of The Eighth National Conference on Artificial Intelligence*, pp. 183–189, 1990.

[6] Shapiro, S.C.: The Turing Test and The Economist. *SIGART BULLETIN*, Vol. 3, No. 4, pp. 10–11, October 1992.

[7] Turing, A.M.: Computing Machinery and Intelligence. *Mind*, Vol. 59, No. 236, pp. 443–460, 1950.

[8] Walker, M.A.: Redundancy in Collaborative Dialogue. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, volume 1, pp. 345–351, 1992.

[9] Weizenbaum, J.: ELIZA–A Computer Program for The Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, Vol. 9, pp. 36–45, 1978.