対話システムの評価における一般的推論能力の要請

橋田浩一
電子技術総合研究所

伝康晴
ATR音声翻訳通信研究所

長尾確
ソニーコンピュータサイエンス研究所

柏岡秀紀
ATR音声翻訳通信研究所

酒井桂一
キヤノン情報メディア研究所

島津明
NTT基礎研究所

1 はじめに

自然言語処理に限らず、人工知能一般における最も重要な課題は、開放性(openness) — 処理すべき情報の範囲が限定されていないこと — への対処である。したがって、そうした技術の評価は、開放性への対処の優劣を評価することによって、この課題に関する研究を促進するものでなければならない。これは、対話システムの評価に関しても同様である。

しかし、開放性に一般的な仕方で対処することは現在の技術では非常に難しいから、評価法の設計には細心の注意を要する。現存の技術ではまったく歯が立たない課題を用いて評価を行なうと、評価の対象となるシステムがほとんどなくなるか、あるいは、課題の十全な達成に結び付かないその場しのぎの対処法が多用されるかであり、いずれにせよ評価の意義が失われてしまうだろう。その場しのぎの方法では対処できないような本格的な開放性を含み、なおかつ現存の技術でもある程度は達成できるような課題を用いる必要がある。評価の対象は公募によって集めることになるだろうが、多くの参加者を得て評価の信頼性を上げるには、参加し易いようにある程度は敷居を低くしておくべきである。ただし言うまでもなく、開放性を含むためには、課題は易し過ぎてもいけない。現在の技術水準に照らしてちょうど良い難易度の課題を用い、技術の進歩に伴って難度を上げて行く必要がある。

また、評価法が広範な社会的認知を得るには、評価の結果がわかりやすいことが重要である。それには特に、評価を行なうための課題が人工的なものではなく、非専門家が日常的な直観で理解できる[1]ものであることが望ましい。開放性を含み、かつ日常的に理解可能な課題は少なくないだろうから、このことは評価法の設計において大した困難にはならないかも知れない。

2 対話リーグ戦

対話リーグ戦(DiaLeague) (橋田, 伝, 長尾, 柏岡, 酒井, ・島津, 1995)は、自然言語対話システムの能力を総合的かつ客観的に評価するために企画されたコンテストであり、参加するシステムの総当たりによるリーグ戦の形をとる。各「対戦」では、2つのシステムの間での対話による何らかの課題の遂行が要求される。課題の達成度に応じて両システムに同一の得点が与えられ、達成度はほぼコミュニケーションの効率に依存してい

[1]伝(1995)はこのことを実世界性(real-worldness)と呼んでいる。

るので、多様な相手と効率よくコミュニケーションできるシステムが全体として良い成績を収めることになる。DiaLeagueが評価しようとしているのは、統語解析や文生成のような要素技術の性能ではなく、対話システムとしての総合性能、すなわち対話に必要なさまざまな情報の統合的処理である。

1995年7月に行なわれたエキシビション・マッチ(DiaLeague'95)と1996年3月に行なわれる第1回本戦(DiaLeague'96)では、**経路課題**(route task)という課題を用いている。これは、2つのシステムに与えられた微妙に異なる2つの鉄道路線図の共通部分に含まれる、スタートからゴールまでの経路を見付ける、という課題である。DiaLegue'95で用いた路線図の例を図1に示す。路線図は実際には記号的なデータとして入力さ

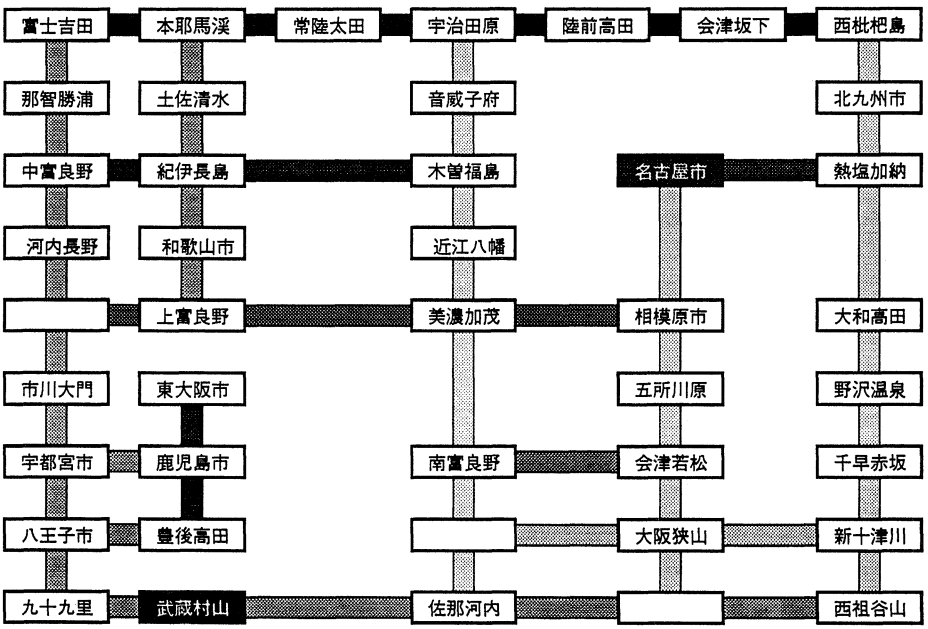

図1: 路線図の例

れる。DiaLeagueの詳細に関しては、http://csl.sony.co.jp/dialeague/ を参照していただきたい。

しかし、佐藤(1995)も指摘するように、これでは問題の構造が簡単であるために単純な対話しか行なわれず、多様な対話能力を正しく評価できないようである。すなわち、経路課題は開放性を欠いていると考えられる。人間同士の自然言語による対話では、量化や条件節などを含む複雑な表現を用いた多様なコミュニケーションが自ずと行われるが、経路課題のような単純な設定の下では、そうした表現を使う必要がないので、計算機同士の対話は易きに流れ、単調なものとなる。つまり、両システムが想定している解候補の集合の共通部分をインクリメンタルに絞り込んで行くことで課題を遂行することができ、それには、「〜まで行けますか」のような質問とそれへの応答だけで基本的には事足りる。たとえば、単なる「はい」とか「いいえ」ではなく「…までしか行けません」のような応答によって情報伝達の効率を上げる、というような工夫もいくつか考えられるが、それもこの課題に特定のその場しのぎの工夫で十分であり、より一般的な対話の方略を考える必要はないだろう。

要するに、問題(意味)が単純である場合には、かなり限られた種類の言語表現と対話パターンによってコミュニケーションを十分達することができる。つまり、これまでの課題は、意味のレベルと表現のレベルが一致したままで対処できる程度の複雑性しか持たなかったので、それでは十分効率的に対処できないように課

題を複雑化するにする必要がある。課題(意味)が複雑になれば、コミュニケーションの効率を向上させるために、量化などを含むマクロスコピックな表現を使ったり、アナロジーなどに基づいて新しい表現関係を動的に作り出したりする必要が生ずるだろう。

3 課題の開放性

現在の経路課題は2つの地図の共通部分に含まれる経路の探索を要請しているが、2つの地図の和に含まれる経路の探索を要請する課題も考えられる。この形の課題についてはDiaLeague実行委員会においても当初から検討されていたが、難し過ぎるという理由で採用されなかった。しかし、適当な条件を加えることによって難度を調節することも可能だろう。その場合の主要な課題は、自分の知っている路線図にない結線がどこにあるかに関する手懸かりをどのようにして与えるかだろう。もしもそのような手懸かりがまったくない場合、すなわち、繋がっていないように見えるどの2つの駅に関してもその間に実は結線がある確率が等しい場合は、虱潰しに結線があるかどうかを聞かざるを得ず、それによって対話に複雑な構造がもたらされるとは思えない。ところが逆にその手懸かりがあからさま過ぎると、路線図の共通部分に含まれる経路を探索する現在の課題と本質的に同じことになってなってしまうだろう。したがって、その中間で適当に折り合う必要がある。

また、ロボットのナビゲーションのようなパターン認識的な側面を導入して経路課題を拡張することも可能だろう。たとえば、一方のシステムがロボットの役割を演じ、他方がロボットに案内の情報を与える、という設定を考えよう。ロボットは自分の周りの環境だけを画像として知覚し[2]、案内役は大雑把な地図を持っていて、対話によって情報を交換しながらロボットの目標への到達を目指す、という課題では、画像と言語の間での情報統合が要請される。「四角柱がある所ではたいてい右折すればよい」のような量化表現によって情報伝達の効率化を図る必要も生ずると予想される。

最も一般的で開放性の高い課題として、たとえば、両システムに2つのよく似た公理系を与え、これらの公理系に共通の証明(あるいは、2つの公理系の和における証明)を求めさせる、というものが考えられる。その一般性は、ほとんどあらゆる知識が公理系で表現できるという事実に基づいており、コンテストに参加するシステムには非常に一般的な推論能力が要請されるはずである。課題としてはやはり、日常的な直観の(マクロスコピックな)レベルで意味が理解できるものが望ましい。そうした意味に基づく比喩や類推を含むさまざまな表現が、対話の効率を高めるために必要となるような課題を設計したい。

必ずしも定理証明課題まで行かなくても、経路課題に新しい意味付けを導入することによって開放性がもたらされる可能性もある。定理証明課題においては、各公理をたとえばリテラルが3個以下のHorn節に限定しても一般性は失われないが、これはその一般性が公理の組合せの構造の多様性に由来しているからである。したがって、経路課題の範囲内でも同様に路線の組合せの多様性によって課題の開放性を高めることができると考えられる。路線のある部分の大雑把な形が文字などの図形になっており、そのような意味の理解に成績が依存するような課題が考えられるだろう。また、課題が十分複雑であれば、ロボットにまつわる課題と同様に、「『田』の付く名前の駅ではたいてい乗り換えればよい」のような量化表現などを使うことによってコミュニケーションの効率が高まることもありうる。もちろん、こうした意味付けや量化の構成法がいくつかの定型的なパターンに限られてしまってしまっては開放性は達成できないから、そこには事実上無限の多様性が必要になる。

人間にやらせてみると、多くの種類の課題の遂行において複雑な対話が自ずと行なわれるが、計算機の場

[2]ロボットと画像はシミュレーションでも構わない。

合には人間の場合よりも対話のパターンが単調になりがちである。特に、人間の対話では確認や問い直しなどが頻繁に行なわれるのに対し、計算機同士の対話ではそのようなことがほとんどない。人間の対話がそのような側面を持つ主な原因は、人間の作業記憶の容量が限られていることである。計算機のプログラムで無理に作業記憶の容量を制限するのは不自然であるし、自由参加のコンテストでそれを強制するのは不可能だろう。しかし、たとえば(路線図などの)課題の設定が時々刻々と変化しており、しかも各時刻にはその一部分しか見られない場合には、古い情報が正しい確率が時間とともに減衰して行くから、計算機にとっても作業記憶が制限されているのと似たようなことになり、確認のための発話などが必要になる。このような時間に関する開放性についてもさらに検討する必要がある。

4 おわりに

本稿では主として対話システムを評価するための課題の一般化/複雑化について考えたが、評価法の設計においては他にも考慮すべき問題が多い。課題の設計に関連して、課題が難しいと評価の対象となるシステムが少なくなってしまうという問題が生ずる。特に定理証明課題では、参加が難しくなり過ぎるかも知れないので、少なくとも初めのうちは、参加し易くするための工夫が要るだろう。また、ロボットの課題による評価に与るには、対話プログラムと(たとえシミュレーションでも)ロボットを含む複雑なシステムを作る必要があり、そのような形で評価に参加するのは非常に難しいだろう。しかし、たとえば対話プログラムだけとか対話とロボットの統合の部分だけに関して評価することは可能である。それには、システムの各モジュールごとに標準的なプログラムを用意しておけばよい。こうした考察と実践を重ねることによってDiaLeagueを洗練して行く予定である。

参考文献

伝康晴 (1995). 計算機同士の対話 —対話リーグ戦とその周辺の話題—. 認知科学会冬のシンポジウム論文集.

橋田浩一, 伝康晴, 長尾確, 柏岡秀樹, 酒井桂一, 島津明 (1995). 対話リーグ戦: 対話システム性能評価コンテストの提案. 『言語処理学会第1回大会年次発表論文集』, pp. 309–312.

佐藤理史 (1995). 対話リーグ戦'95用プログラムの作成—対話からの知識獲得—. テクニカル・レポート IS-RR-95-18I, 北陸先端大科学技術大学院大学. ISSN 0918-7553.