対話リーグ戦: 対話システム性能評価コンテストの提案

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

自然言語処理システムの性能評価に関してはすでにいくつかの試みかあるが、対話のような総合的な能力を要求される技術に関しては、客観的でしかも科学技術の進歩を促すような評価法が確立していない。本稿では、自然言語処理研究の発展を促進するような仕方で対話システムの総合性能を客観的に評価する方式を提案する。基本的アイデアは、対話を通じて2つのシステムに何らかの課題を行なわせるというものであり、その課題をうまく選択することによって、通常の意味での対話の能力を適確に評価することを目指す。各「対戦」においては作業の達成度等に応じて両方のシステムに同じ得点を与えることにより、自然に協調的な対話が行なわれるようになる。また、コンテストに参加するシステムのすべてのペアにわたって対戦を行なう対話リーグ戦 (**DiaLeague** と呼ぶ) の形を取ることで、人間の介在が不要になるとともに、対戦相手の多様性に応じた幅広い対話能力を要求できると考えられる。

2 評価法の問題

技術が十分に成熟していない段階で不用意な評価を行なうと、確実に成績を上げるために専ら無難な既存技術が使われ、却って研究の進展を阻害してしまう結果になりがちである。たとえば、DARPA が1987年以来ほぼ毎年行なっている、計算機プログラムによるメッセージ理解の競技会 (Message Understanding Conference; MUC) においては、新聞記事などの文章から予め定められた種類の情報 (企業の合併等に関する文章の場合には、何という会社と何という会社がどういう条件で合併するのか、など) を抽出するという課題が用いられるが、この課題では処理すべき情報の範囲がかなり明確に限定されているため、キーワードスポッティングやテンプレート照合などの使い古された技法ばかりが多用される結果を招いてしまっており、談話理解における真に難しい問題を解決するための研究が正当に評価されているとは言い難い。また、Turing テストを模した Loebner 賞 [3, 4] では、人間と見分けが付かないような対話を行なうという、明らかに現在の技術水準を越えた課題に対処するまともな方法がないので、人間のタイピングのリズムやタイプミスなどをまねるなどの小手先のテクニックを使ったシステムが良い成績をおさめており、これも先進的な研究の努力を促すコンテストにはなっていない [4] 。

こうした例を見ると、自然言語処理のような発展途上の分野で、文章理解や対話処理のような総合的な能力を要求される技術に関して、研究の発展に寄与するような評価を行なうには、システムの理論的基盤などに立ち入って主観的に評価するとか、システムの能力を複数個の基本的な側面に分けて別々に評価するとかいった方法が必要であるように思われる。しかし、基盤となる理論に関する評価は主観的なものにならざるを得ない。また、システムの能力を基本的な側面に分けてしまっては総合的能力の評価にはならない。統語解析や文生成を何らかの精度に関して評価した結果、すべての面でシステム A の方がシステム B より優れていたとしても、そうした部分的な能力を組合せた総合的な能力においてはシステム B の方が良い、というようなことがあり得るだろう。重要なのはもちろん総合的能力の方であり、以下で考えたいのはこれを客観的に評価する方法である。

3 対話システムの評価

対話システムを総合的かつ客観的に評価するために、何らかの課題を対話を通じて行なわせることを考える。通常の意味での対話能力を正しく評価するには、その課題は、**相互作用主導的** (interaction-oriented) かつ**協調誘導的** (cooperation-inducing) かつ**開放的** (open) でなくてはならないだろう。相互作用主導的とは、主に考えるだけでできる類の課題ではなく、他者からインタラクティブに情報を得ることが作業の大部分を占めるような課題だという意味である。協調誘導的とは、その相互作用を協調的に行なった方が成績が良くなるということである。開放的であるとは、処理すべき情報の範囲が開かれており、狭く限定できないという意味である[1]。

相互作用主導性は、効率のよい情報伝達という限定された意味での対話能力をなるべく純粋に評価するための条件である。Loebner 賞の場合には、人間らしい対話なるものの定義が明確でないために主催者が意図しなかった小手先のテクニックがまかり通ってしまったが、情報伝達の効率と正確さが成績に正しく反映されるような課題を用いることによってその問題は回避できる。その場合、対話の人間らしさとか自然さのような側面が度外視されてしまう可能性はあるが、対話システムの応用における重要性を考えればそうした側面に関する評価方法はさしあたって未解決問題としておいてよかろう。

協調誘導性は、誠実な対話が行なわれることを保証するための条件である。自然言語による対話に限らず、一般にコミュニケーションにおいて、直接的な発語内行為 (illocutionary act) のレベルでは協調性が成立する[2]が、対話が嘘や誤解を招く表現を含み、全体としては協調的でないこともありうる。しかし、対話システムの応用において第一義的に求められるのは通常の協調的な対話の能力と考えられるから、評価用の課題において非協調的な対話が行なわれることは望ましくない。すなわち、課題は協調誘導的でなくてはならない。もちろん、嘘やはったりまで含めた対話能力が応用上有用である可能性もあり、将来はそこまで考慮した評価法が必要となるかも知れないが、現在の技術水準では、対話を誠実で協調的なものに限定し、評価の対象を明確にした方がよいだろう。

相互作用主導性も協調誘導性も、開放性に反しない。つまり、情報伝達という明確な目的を持った協調的な対話というだけでは、MUC の場合と異なり、処理すべき情報の範囲を過剰に制限することにはならない。ただし、そこで広い範囲の情報を処理させるためには、多様な対話を行なせる必要があり、それには評価方法に何らかの工夫を要する。人間が対話システムの対話の相手を務めることによって、人間に合わせて多様な対話を行なわせる、という方法も考えられるが、人間の方がコンピュータに合わせて少数の簡単な表現しか使わなくなるかも知れない。また、特に評価すべきシステムが多い場合、人間の側の疲労や慣れや個人差によって、評価の公正を期し難いだろう。

そこで、参加する対話システムの全ペア (自分自身とのペアも含む) についてシステム同士で「対戦」して総合成績を競うリーグ戦方式を考え[3]、これを**対話リーグ戦** (Dialeague) と呼ぶ。この評価法の下では、多くの相手との対戦で総合的に良い成績を上げるために多様な対話能力が求められることになる。また、能力の高いシステムは特に自分自身との対戦で高得点が取れるので、研究開発の努力が報われる。さらに、人手の介在を最小限にして公正を図りつつコストを削減することもできる。このようにシステム同士が対戦する場合、協調誘導性を満たすには、各対戦において両方のシステムに課題の達成の度合に応じた同一の得点を与えればよい。

ただし、システム同士のリーグ戦によって、システム間の比較による相対的評価はできるが、絶対的評価はできない。これに関連して、相手が人間である場合の対話能力を評価することも不可能である。人間相手の絶対的な対話能力を適正に評価することには当然意義があると考えられるが、これに関しては今後の検討事項としたい。

[1] 相互作用主導性と協調誘導性は評価の対象が対話能力であることに依存しているが、開放性は人工知能システム一般の評価に広く妥当する規準である。

[2] つまり、自分の意図する意味内容が受信者に正しく伝わることをメッセージの送信者が望むのはもちろん、たとえ受信者が送信者の誠実性を疑っていても、送信者が表面上意味している内容を理解することは嘘を見破る上で有用だから、受信者も送信者が意図した意味内容を理解することを望む。ちなみに、こうした協調性は自然言語の構造にも反映されている [2]。

[3] 参加するシステムが多過ぎて総当たりが不可能なら、予戦リーグと決勝リーグに分ければよい。ただし、後述のようにリーグ戦の結果は相対的評価なので、予戦リーグの成績を決勝リーグの成績に持ち越してはならない。

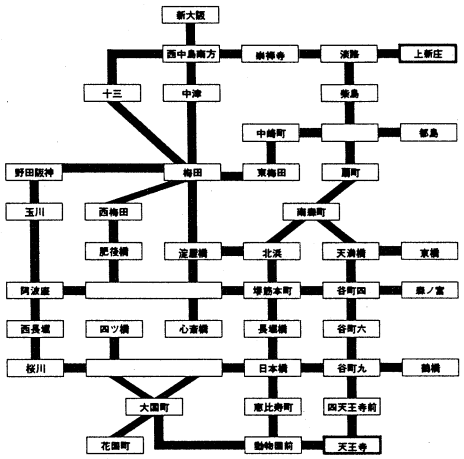

図 1: 一方のシステムに与えられる路線図

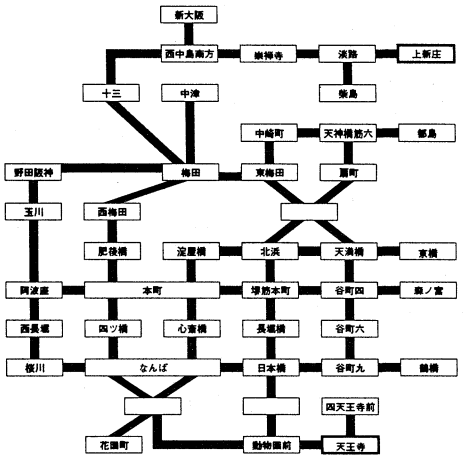

図 2: 他方のシステムに与えられる路線図

4 経路課題

DiaLeague の各対戦で用いる課題として現在われわれが考えているのは、**経路課題** (route task) である。これは地図課題 (map task) [1] にヒントを得たものだが、現状の技術で実現可能な対話システムによって遂行可能で、なおかつ相互作用主導的となるように考慮されている。

経路課題とは、たとえば図 1と図 2のような鉄道の路線図を与えられた 2 つのシステムが、指定された出発駅と到着駅を結ぶ共通の経路を対話によって見付けるというものである。2 つの路線図は図 3に示した原図の不完全な複製である。出発駅と到着駅は 2 つの路線図で一致しているが、原図中に×で示されている切断箇所はどちらか一方の路線図だけで繋がっており、また 2 つの路線図でいくつかの駅名が消えている。さしあたり、パターン認識処理の必要をなくすため、路線図は実際には画像ではなく記号的な表現としてシステムに与える。また、図の例題は実在の路線に基づいているが、実際の DiaLeague では架空の路線図を用いる。

2 つのシステムは、切断箇所やわからない駅名に関する情報を対話によって獲得しながら、両者が共通に利用できる経路を制限時間内に見付けなければならない。両システムは、見付けた経路の短かさ (駅の個数の少なさ) と対話の短かさ (文字数の少なさ) に応じた同一の得点を与えられる[4]。例題では、出発駅は「天王寺」、到着駅は「上新庄」であり、「天王寺→動物園前→恵比寿町→日本橋→長堀橋→堺筋本町→本町→肥後橋→西梅田→梅田→十三→西中島南方→崇禅寺→淡路→上新庄」という経路が長さ 15 で最短となる。

経路課題においては、最初に一方のシステムが自分の路線図の全情報を他方に教え、教えられたシステムが正しい最短経路を求め、これを相手に教えるというふうにすれば、対話らしい対話をせずに課題が達成できる。しかし、2 つの路線図はほとんど同じものだから、この方法は両方の間の差に関する情報だけを対話によって漸進的に伝え合う方法よりも著しく効率が悪い。たとえば、最短経路に関する両システムの仮説を対話によって擦り合わせて行って一致させる方法を用いれば、路線図に関する情報のほとんどは発話しなくても済むだろう[5]。すなわち、経路課題は相互作用主導的である。ほぼ上記と同じ例題を用いて人間の被験者に経路課題を解かせる実験をいくつか行なってみた結果、課題の遂行には 5 ～ 6 分を要し、照応表現や確認などのさまざまな要素を含む対話がなされた。対話システムの場合にも同様のことが起こると期待される。

対話はすべて日本語で行なうものとするが、使用できる語彙や文型などは一切規定しない。2 つのシステムは、まずそれぞれの路線図を入力した後、主催者が用意した監督プログラムを介して対話を行なう。監督プログラム

[4]経路が誤っていたり 2 つのシステムの間で一致しなかったりした場合、あるいは制限時間内に経路を見付けられなかった場合にも部分点を与えるのがよかろう。

[5]正確には、そうであるかどうかは路線図によるが、そうなるように問題を作ることは可能である。

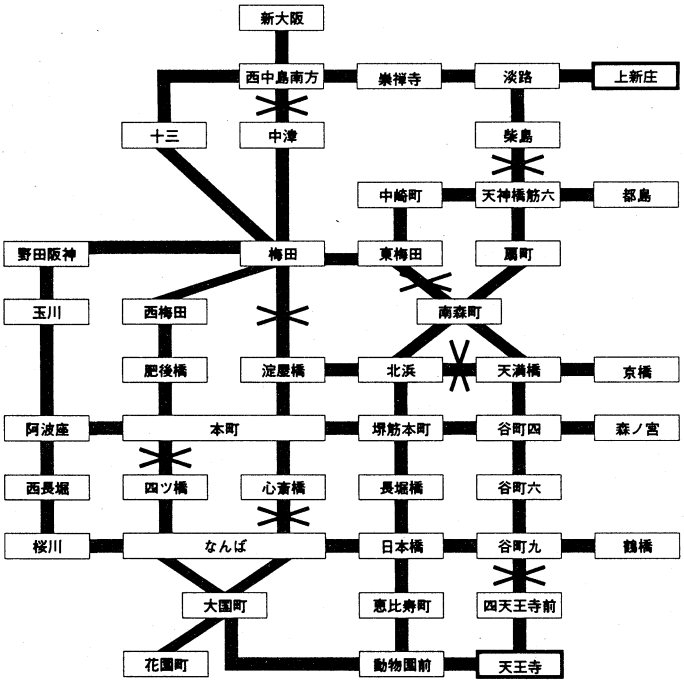

図 3: 2 つの路線図の原図

は、各システムから送られてくる発話データを読み、それをモニタしながら、他方のシステムに転送する。今の所、2 つのシステムの発話は簡単のため時間的に重ならないことにしてあるが、より自由度の高い対話を可能にするため、この制限は撤廃すべきだろう。

5 おわりに

対話システムの総合的な対話能力を評価する方法を提案した。この提案に基づいて実際に対話リーグ戦を行ないたい。1995 年中にエキシビション・マッチを行なった後、1996 年に第 1 回コンテストを開催する予定である。参加を希望される方は著者まで連絡をいただきたい。

参考文献

[1] A. H. Anderson, et al. The HCRC map task corpus. *Speech and Language*, 34(4):351–366, 1991.

[2] Koiti Hasida, Katashi Nagao, and Takashi Miyata. A game-theoretic account of cooperation in communication. Technical Report TR-95-6, Electrotechnical Laboratory, 1995.

[3] 小谷 善行・松原 仁・ 大沢 英一. コンピュータは人間に勝てるか?! 情報処理, 34(3):275–284, 1993.

[4] Stuart M. Shieber. Lessons from a restricted turing test. Technical Report TR-19-92, The Center for Research in Computing Technology, Harvard University, 1994.