

概念ベースを用いたネットワーク型 「言葉遊び」の提案

松澤 和光[†] 阿部 明典[†] 金杉 友子^{††} 永森 千晴^{††}

[†]NTT コミュニケーション科学研究所 ^{††}NTT アドバンステクノロジー(株)

1 はじめに

当研究所では、与えられた知識が不完全でも、これを常識で補って概略的な答えを導く「アバウト推論」の研究 [3, 4] を進めている。アバウト推論は、推論する対象を「単なる記号」ではなく、常識で補間可能な「意味を持ったもの」として扱うことを特徴とする。このため、システム外部からの入力と言葉のままとし、これを解釈するために「語の意味」(概念)を体系化した「概念ベース」をシステム内に持つ構成とした。

この概念ベースについては、電子化文書等から完全に機械的に構築する手法を開発し [2]、既に4万語規模のものを実際に構築済みである。しかし、構築された概念ベースに人間が持つと同様な「語の意味」が本当に蓄積されているか、その評価は難しい。

著者らは、この概念ベース評価法の一つとして、人間が行なう「言葉遊び」を概念ベースを使って自動的に行なう方式 [1] を先に提案した。対象とする言葉遊びは、連想しりとり、クロスワードパズル、などなど等であり、概念ベースの関連語検索機能を用いて作成した解答を人間が見て、「人間の感覚と一致しているか」評価した。しかしながら、こうした感覚的な評価法は、評価者に依存する部分が多く、客観的な評価を行なうためには、評価対象を広く一般の人々に拡大する必要がある。

一方、こうした「言葉遊び」への自動解答システムは、他にはほとんど類例がなく、新種のエンタテインメントあるいはイベントとしての可能性を秘めている。これを商業的に実現するためにも、システムを多くの人々に提供する仕組みが重要となる。

しかしながら、これら「言葉遊び」のアプリケーションに用いる概念ベースは、4万語規模程度の語彙に対してでさえ数十MB程の容量を要し、アプリケーションをさらに拡張した場合にはより多くの容量が必要となる。こうした膨大なデータを、複数のコンピュータ各々に置いたり、種々の媒体で配布するのは非効率的である。

ところで、Mosaic や Netscape の出現により、ネットワーク上での情報閲覧、商業取引等のアプリケーションが盛んに行なわれている。これらはホストコンピュータに、実行に必要な全てのファイルを置き、それをネットワーク経由でアクセスすることで、実現している。このようなネットワークを介してのアプリケーションの実行は資源の分散の意味もあるが、巨大な資源を一箇所で管理出来、それをCD-ROM等のメディアで配布する必要がないという意味でも重要である。

そこで、今回、概念ベースの評価対象と適用範囲の拡大に向けた検討の第一段階として、「言葉遊び」をネットワークを通じたアクセスに対応できるように拡張した。この実験システムについて報告する。

2 概念ベースを用いた「言葉遊び」

まず、概念ベース技術と、これを用いた言葉遊び— などなど、および連想しりとり— について簡単に紹介する。詳しくは文献 [2, 1] を参照されたい。

2.1 概念ベース技術

概念ベースでは、 N 個の語の各々の意味(概念 g_i)を、同じく N 個の語から選んだ g の特徴を表す語(属性 p)、および p の重要度 q の対の集合で表す。すなわち

$$\text{概念 } g_i = \{(p1, q1), (p2, q2) \dots\} \quad (1)$$

ここで各 p, q は、国語辞典等における見出し g に対する語義文から、自立語とその出現頻度として機械的に抽出する。例えば語「馬」について「首と尾に毛が生えた動物で、走るのが早い動物・・・」云々との語義が書かれていた場合、

$$\text{「馬」} = \{(\text{首}, 1), (\text{動物}, 2) \dots\} \quad (2)$$

等となる。

もちろん、単に語義文から自立語を抽出しただけでは真に「その語の特徴を表す」属性は十分には獲得できない。そこで、まずこのようにラフな概念ベースを作成した後で、概念ベース自身を参照しながら属性データの質の向上をはかる技術—自己精練化技術を開発した(詳しくは文献[2])。

この結果、概念ベースは $N=4$ 万の語を各々 N 次元の属性ベクトルとして表現した $N \times N$ の行列として構築される。これを用いて、ある語 a に対し a に関連した語 $K(a)$ を、語 a のベクトルが含む属性 p 、語 a を属性 p として含む語、語 a と似た属性ベクトルを持つ語 c 、等として検索することができる。また、この際に重要度や類似の度合を「関連度」として計算することができる。

2.2 言葉遊びの実現方式

(1) なぞなぞ

ここでは「クリはクリでも口から出るクリは?」といった「語 a は a でも語集合 C に関連する a は?」の形式の問題を扱う。これは

- 単語 a に音韻的に近い語
- 語集合 C に関連した語

の2条件を考えて解く問題と考えられる。

そこで、まず語 a 「クリ」がそのまま、または音韻的に近い形(「グリ」等)で含まれる単語群 A を解候補群として選出する。次に、概念ベースを用いて語集合 C (口, 出る)の語の関連語検索を施す。 A 中で関連語として検索され、関連度がもっとも大きい語を正解とする。

この例では、「あまぐり」、「しゃっくり」、「くり返し」...等の候補から、正解である「しゃっくり」が選出されるわけである。

(2) 連想しりとり

連想しりとりは、TVの番組の中でも行なわれているゲームの一つで、例えば「バナナといたら黄色」のように、ある語から連想する語を、「語 a といたら語 b 」の形で続けて回答していくものである。そこで、概念ベースを用い、与えられた語 a の関連語を検索して、そこから無作為に選出した1語を語 b として回答を行なう方式とする。

3 言葉遊びのネットワーク化

今回、ネットワーク化したのは、「なぞなぞ」と「連想しりとり」である。これらのアプリケーションは、ユーザが質問をしてそれに対する答をコンピュータが返すという形式で行なわれている。従って、フォーム形式でメッセージをホストコンピュータに送って、ホストコンピュータで、CGI (Common Gateway Interface) を実行をして結果を画面に表示するという形式を採った。表示の形式は他の面をクリアしないという点から、フレームを用いている。新しいブラウザを開くという手法もあるが、一つのブラウザで表示出来るフレームを採用した。更に、後で詳述するが、このフレーム形式の良い点は同じ状態で何回も様々な質問を続けられる、履歴を辿っても他のフレームには影響を与えないという点等である。

(1) なぞなぞ

なぞなぞの例題は小児向けクイズ本から「語 a は a でも語集合 C に関連する a は?」の形式の問題を抜き出して利用した。これらの問題には番号がふってあって、番号を入力することで、その番号の問題に対する答の候補を返すというシステムである。質問に対する答は、質問に言葉の一部が含まれているので、それを含む語のうち、関連度の高いものに高い尤度を与えて表示する。

ネットワーク化に際しては、質問を解く部分を CGI 化し、ユーザからネットワークを介して送られてきた問題を解き、フレームに表示するという形で行なう。左側のフレームに問題を表示して、右下の入力フレームに解きたい問題の数字を入れ、それを送ることにより、右上のフレームに結果を表示している(図1)。正解が得られた場合は、正解したという旨を示し(図1)、正解が得られなかった場合は、正解のある位置を示す(図2)。画面をフレームで分割しているので、右下の入力枠に数字を入れて送るだけで、続けて他の問題を解くのも可能である。更に、フレーム内の back、forward により、前に出した答を見直すことも容易である。



図 1: 「なぞなぞ」の画面(正解の場合)

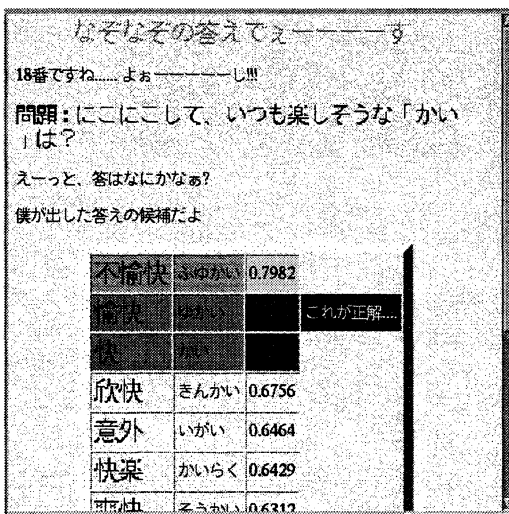


図 2: 「なぞなぞ」の画面(不正解の場合)

(2) 連想しりとり

ある言葉に対して、連想出来る言葉を順番に出していくゲームであり、制約として、前出の言葉は当然、駄目、いいかえだけは駄目等というのがある。本システムでも与えられた言葉に対して、関連度の高い概念を表示するというを同様の制約の下で

行なう。

ネットワーク化に際しては、前記「なぞなぞ」と同様に、質問を解く部分を CGI 化し、ユーザからネットワークを介して送られてきた問題を解き、フレームに表示するという形で行なう。但し、コマンドライン版では、表示した答に対して、可否を聞き、否の場合は、別の答を用意するということを行なっているが、ネットワーク版では、ネットワークを介しての処理を考え、ホストコンピュータが最初の答に対して答を出し、その答に対する連想語を順々に出すという形式を採用した。

本システムでは殆んどの処理は右側のフレームで行なわれ、右下のフレームに最初概念、連続連想したい語の最大数を入力し、それを送ることにより、右上のフレームに解答を表示する(図 3)。左側のフレームは単にインタフェースのためだけである。図に示したものは、「蛸」を最初の言葉として入れた例である。画面では、9 語しか表示されていないが、実際は、100 語表示されている。なかなか面白い結果が得られていると思う。なお、100 語を出すのに要した時間はホストコンピュータ、クライアント共に、SS20 151 で約 2 分(実際の時間、CPU time ではない)なので、少し時間がかかると感じる程度である。

4 あとがき

語の意味を体系化した概念ベースを用いて、コンピュータによる「言葉遊び」をネットワークを介して行なう実験システムを紹介した。本システムの実現により、概念ベースが人間の持つ「語の意味」を反映しているか否かの評価実験を大規模に行なえる見通しが得られただけでなく、こうした「言葉遊び」を具体的アプリケーションとして実用化する可能性も開かれたと考える。

なお、今回は言葉遊びとして「なぞなぞ」と「連想しりとり」のネットワーク化を行なったが、これらのさらなる拡張や、先に提案した「クロスワードパズル」「なぞかけ」等についても、順次、ネットワーク化を図っていく計画である。

さらに、これら以外にも概念ベースを用いた種々の言葉遊びが考えられるだろう。あるいはネットワークを積極的に利用して、今までになかったような言葉遊びを生み出せる可能性もある。これらは商



図 3: 「連想しりとり」の画面

業的な価値ばかりでなく、きたるべきネットワーク社会における「新しい言葉文化」を切り開き、ひいては実り豊かなコミュニケーション社会の実現に寄与できるのではないだろうか。

そんな夢を描きつつ、上記のような「新・言葉遊び」のアイデア・コンテスト的なものを企画中である。まだ詳細は決まっていないが、早ければ3月28日の本報告のポスターおよびデモ展示において発表できるので、ご期待頂きたい。あるいは、当日までに開設する予定のホームページ

(<http://www.kecl.ntt.co.jp/banana/>)

を御覧あれ。

参考文献

- [1] 金杉, 松澤, 笠原: アバウト推論の「言葉遊び」への適用, 信学技報, *NLC96-31*, pp. 1-7 (1996)
- [2] 笠原, 藤本, 松澤, 石川: 精錬化に基づく概念ベース構成法, 信学技報, *DE95-7*, pp. 49-56 (1995)
- [3] 松澤, 石川, 湯川, 河岡: アバウト推論 — 常識的な推論を目指して —, 人工知能学会研究会資料, *SIG-FAI-9401-1*, pp. 1-8 (1994)
- [4] 松澤, 湯川, 笠原, 藤本: アバウト推論技術 — 常識を持ったコンピュータの実現を目指して —, *NTT R&D*, Vol. 45, No. 11. pp. 115-120 (1996)