

E-070

コンピュータ会話における非常識文の抽出及びその応答方式

Extraction of Illogical Discourse and the Reply Method in Computer Conversation

倉永 久仁彦†
Kunihiko Kuranaga

吉村 枝里子†
Eriko Yoshimura

渡部 広一†
Hirokazu Watabe

河岡 司†
Tsukasa Kawaoka

1. はじめに

現在コンピュータの普及に伴い、小さな子供から老人まで様々な年代でコンピュータが利用されている。しかし入力装置は、コンピュータができた当初からほとんど変わらずマウスやキーボードのままである。それらの入力装置は、小さな子供にとって文字を理解できないためキーボードを打つのが困難な場合があり、老人も視力が衰えてキーボードの表示文字が読みにくいなどの問題が考えられる。そこで、小さな子供から老人まで行えるコミュニケーションである会話を用いて、コンピュータに人間の意思を伝えるシステムがあれば、今までコンピュータを使用することを敬遠していた人々にとっても、利用しやすい身近なものになる。

現在コンピュータと交わされる会話研究は、人間が話す会話文に対して機械的に応答文を返すだけで、いまだ人間同士の会話のような円滑な会話は実現できていない。

そこで、人間同士の円滑な会話を実現するために、会話を理解し、人間の考える常識文であるかを判断するシステムを提案する。このシステムを利用することで、今までコンピュータと交わされる会話文が全て常識文だという前提で会話をしていたが、非常識文が入力された場合にも正しい会話を行うように人間に促す応答処理が実現できる。これらの非常識文抽出メカニズムと応答システムの実現により、人間とコンピュータとの円滑な会話の実現を目指している。

2. 研究概要

コンピュータによる人間との自然な会話の実現のためには、人間の非常識な会話に対して、システムが非常識だと理解し適切に返答する必要がある。具体的には、人間が「空を歩いたよ」とコンピュータに発話した場合に、システムが「空を歩く」は非常識だと理解し、「空を歩いたのですか?」と問い返せることが必要である。このような応答ができれば、コンピュータが正しく理解をしていないのであれば「空を歩いたような体験をした」などのように説明を補足することができると考えられ、途切れず円滑なコミュニケーションを続けることが可能となる。

また、非常識文には主語や目的語に付属している助詞が誤っており、文の意味を正確に解析できないという問題がある。例えば「図書館は出発する」という入力文を機械的に解析すると「図書館」が主語だと判断するが、人間は「図書館」を目的語として理解し助詞を訂正して「『図書館に出発する』ではないですか?」と発話者に聞きなおす

と考えられる。そこで、人間と同様にシステムでも文の意味が解析できない際に適切な助詞を見つけ、システムの解釈した例を示す応答を行い、発話者に確認をとる処理ができるようにする。

3. 判断知識

非常識文の抽出に利用したシステムについて説明する。

3.1 Webから自動構築された大規模格フレーム^[1]

Web から自動構築された大規模格フレームは、用言とそれに関係する名詞を用言の用法ごとに整理したデータベースであり、Web 上の約 5 億文の日本語テキストから自動的に構築され、約 5 万語の用言からなる。

そのデータベースを利用し、用言と特定の名詞と助詞を指定して日本語テキスト中で何件用いられるか(格フレームヒット件数)検索をかけることができる格フレーム検索システムを使用する。

格フレームヒット件数は、日常的にどれくらいの頻度で用いられるかの指標となる数字であるとし、0 件であれば日常的に用いられない表現であり、数字が高くなればなるほど、頻繁に用いられる表現であるとする。

3.2 シソーラス^[2]

シソーラスには、一般名詞の意味的用法を表したデータベースと用言の文型パターンを示したデータベースがある。非常識文抽出に利用したのは、一般名詞の意味的用法を整理したデータベースで、一般名詞の意味的用法を表す 2700 の意味属性(ノード)の上位下位関係、全体部分関係が木構造であり、約 13 万語(リーフ)が登録されている。

4. 非常識文抽出システム

非常識文の抽出に、NTTコーパス^[2]を利用する方法が考えられる。NTTコーパスは、6000 語の用言から構成される 14800 もの文型パターンを整理したデータベースであるが、Web から自動構築された大規模格フレームと比較すると、人手で構築されているので、文型パターンごとに細かい補足情報が得られるメリットがある。しかし、非常識文抽出に必要な情報は、その文型パターンが用いられるか用いられないかだけであるので、Web から自動構築された大規模格フレームで十分対応できる。さらに、Web から自動構築された大規模格フレームは約 5 億文の日本語テキストから構築されているので、広範囲の文に対し対応でき、Web から自動構築された大規模格フレームを利用の方が適切であると考えられる。

また、非常識文の抽出を行うには、ある言葉が日常生活でどれくらいの頻度で用いられるかの指標となるものが必要である。本稿では、Web から自動構築された大規模格フレームを利用するが、Google^[3]などを用いてWeb検索エン

†同志社大学大学院工学研究科 Graduate School of Engineering, Doshisha University

ジンにかけ獲得した検索件数を、日常的に用いられる指標として非常識文抽出を行う方法も考えられる。しかし、その方法では、システムを使用する度にWebに接続するので非常に時間がかかるという問題点があり、すでにデータベースが構築されていて、非常識文の抽出を行う度にWebに接続する必要のないWebから自動構築された大規模格フレームを利用するのが適切だと考えた。

4.1 非常識文抽出方法

Web から自動構築された大規模格フレームを利用する非常識文抽出システムは図1の流れで行う。

まず会話文に対し形態素解析を行い、主語、目的語、用言のフレームに分ける。「子供が電車を植える」であれば、主語「子供」、目的語「電車」、用言「植える」に分割する。

次に「主語」と「用言」を格フレーム検索システムにかけ、格フレームヒット件数が0件ならば非常識文と判断するが、それ以外の場合は次のステップに進む。例文の場合には「子供」、「植える」、助詞「が」で格フレームにかけ、格フレームヒット件数が16件なので、次のステップに進む。

次に「目的語」と「用言」を格フレーム検索システムにかけ、「主語」の場合と同様に格フレームヒット件数が0件ならば非常識文と判断し、それ以外の場合には常識文と判断するようにした。例文では「電車」、「植える」、助詞「を」で格フレーム検索システムにかけ格フレームヒット件数が0件なので「子供が電車を植える」は非常識文だと判断できる。

但し「主語」や「目的語」に形容詞や形容動詞が付属していた場合にも、形容詞や形容動詞を省いた状態の名詞で格フレーム検索システムにかけ処理を行う。

また Web から自動構築された大規模格フレームには「旅行する」、「運動する」などのサ変接続動詞はデータベースに存在しないという問題があったが、判断できるように対応した。

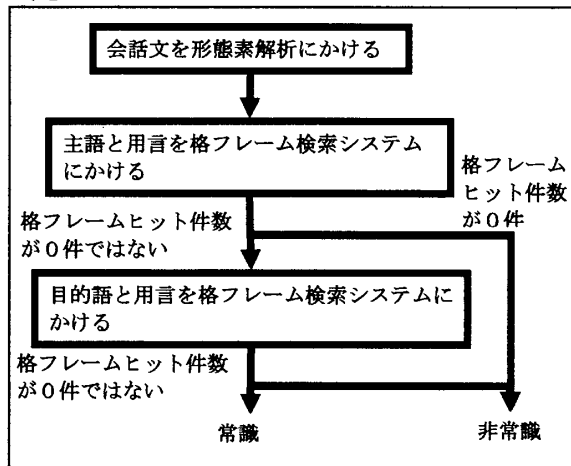


図1 非常識文抽出の流れ

4.2 曖昧文の設定

曖昧文設定の経緯を説明する。

4.2.1 常識文か非常識文かの二極判断の弊害

コンピュータに人間の持つ常識を持たせるという研究目的に沿うためには、ある特定の人々の常識と適合していれば「正解」と判断すればいいのかという問題がある。仮に、ある特定の人々の常識を正解とすると、その人の性別、年齢、

職業など様々な条件の設定をどうすればよいか決定できない。そこで、判断する人を複数人用意し、共通して常識と判断する文のみを「常識文」、共通して非常識文と判断する文のみを「非常識文」と判断を行うことにし、人によって常識か非常識かの判断が分かれる文に対しては、「曖昧文」という評価分類を用意することにする。

4.2.2 曖昧文設定のための実験

曖昧文を設定するために、実験を行った。

アンケートにより評価データを収集し、格フレームヒット件数80以下の評価データ70文を用意する。さらにアンケートで、評価データが常識文か非常識文、若しくは曖昧文であるかを10人に調査を行った。各評価データに対して、10人又は9人が常識文と判断すれば「常識文」、10人又は9人が非常識文と判断すれば「非常識文」、それ以外を「曖昧文」とそれぞれ定義した。

結果を図2に示す。図2より、曖昧文は格フレームヒット件数40以下にしか存在しないことがわかったので、曖昧文と常識文、非常識文の間に設定する閾値は暫定的に格フレームヒット件数40とする。但し、格フレームヒット件数が40件以下の文であっても、常識文、非常識文が存在するので、曖昧文に対してはシソーラスを用いる処理をさらに加えた。

シソーラスを用いる処理について具体的に説明すると、「針を食べる」は格フレームヒット件数が5件で改良前には曖昧文と判断されていたが、「針」のシソーラスの直接の親ノード「刺針」、「作業具」を取り出し、それらと「食べる」で格フレーム検索システムにかけること、格フレームヒット件数がそれぞれ0件で正しく非常識文だと判断できるようになる。

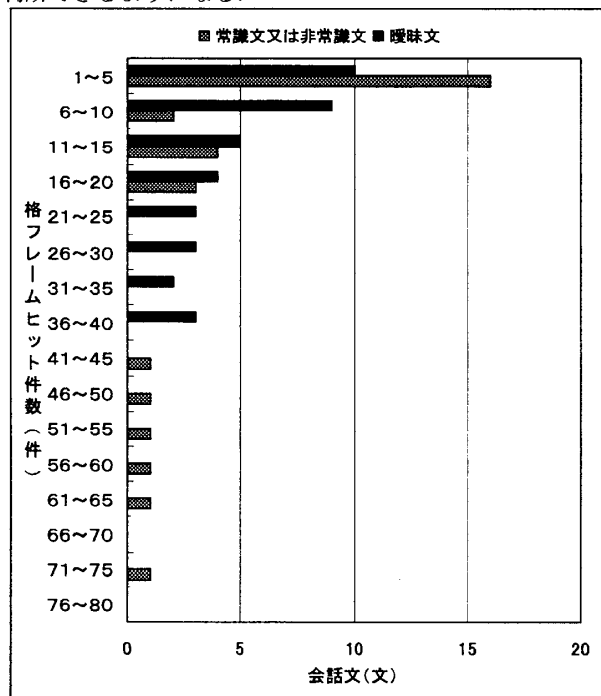


図2 格フレームヒット件数80以下の評価データ

4.3 非常識文抽出システムの評価方法

評価データとして、「主語+動詞」、「目的語+動詞」、「主語+目的語+動詞」の3種類の文章をアンケートにより収集する。また、収集した評価データに対して、評価データが常識文、又は非常識文であるかどうかを

10 人に対して調査を行った。各評価データに対して、10 人又は 9 人が常識文と判断すれば「常識文」、10 人又は 9 人が非常識文と判断すれば「非常識文」、それ以外を「曖昧文」とそれぞれ定義し、常識文、非常識文を各 100 文、曖昧文を 50 文の内容で各 250 文、計 750 文を用意した。また、非常識文抽出システムかけた分類とアンケートにより収集した分類とが一致している場合に正解とした。

4.4 常識文の評価

図 3 に常識文の評価結果を示す。常識文を正しく常識文だと判断できたのは主語 91%、目的語 89%、主語と目的語 84%だった。成功例に「サルが木から落ちる」、「膝が笑う」、「改札をくぐる」などがあった。失敗例としては、「賽を投げる」や「落語を語る」のような日常会話ではあまり用いられないような文が曖昧文と判断された。

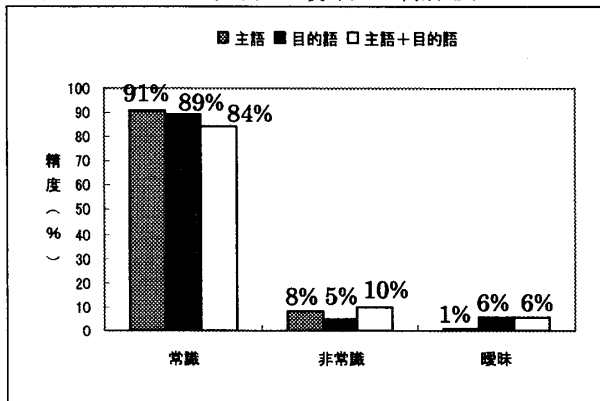


図 3 常識文の評価結果

4.5 非常識文の評価

図 4 に非常識文の評価結果を示す。

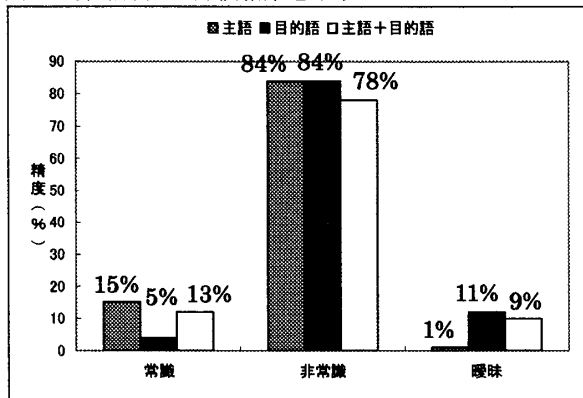


図 4 非常識文の評価結果

非常識文を正しく非常識文だと判断できたのは主語 84%、目的語 84%、主語と目的語 78%だった。成功例として「マイクが歌う」、「蟻が象を踏んだ」、「山で泳いだ」などがあった。失敗例として「山が笑う」は Web 上で比喩表現として用いられていたため常識文と判断し、「眼鏡が飛んだ」は「殴られて眼鏡が飛んでしまった」などのような文が Web 上に存在し、常識文として判断した。

他の失敗例に「父がお年玉を貰う」があった。その原因は、主語と目的語の関係に対して判断を行っていないことだった。つまり、現状のシステムでは「父がお年玉を貰う」に対して「父」が「貰う」と「お年玉」を「貰う」で判断を行い常識文として判断するが、「父」が「貰う」ものとして「お年玉」であることが非常識であるという判断ができるように処理を加えていく必要がある。

4.6 曖昧文の評価

図 5 に曖昧文の評価結果を示す。

曖昧文を正しく曖昧文だと判断できたのは主語 46%、目的語 54%、主語と目的語 36%で、他の常識文や非常識文の再評価結果に比べても精度は高くなかった。成功例としては「庭で寝る」で、「庭」ではなく「寝室で寝る」が適切な文だと考え、非常識文と捉える人もいるし、「庭で寝ようと思えば寝られる」と考え常識文と捉える人も存在するので曖昧文といえ、システムも曖昧文と評価できていた。失敗例として、「トイレで寝る」という例があった。

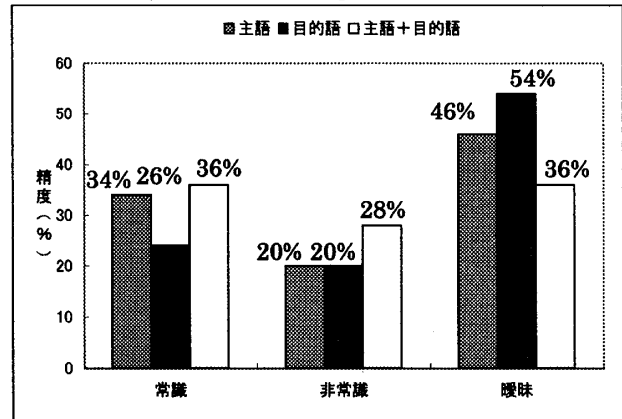


図 5 曖昧文の評価結果

5. 助詞訂正を伴う応答

応答方式について以下で述べる。

5.1 応答処理の流れ

図 6 に応答処理の流れを示す。

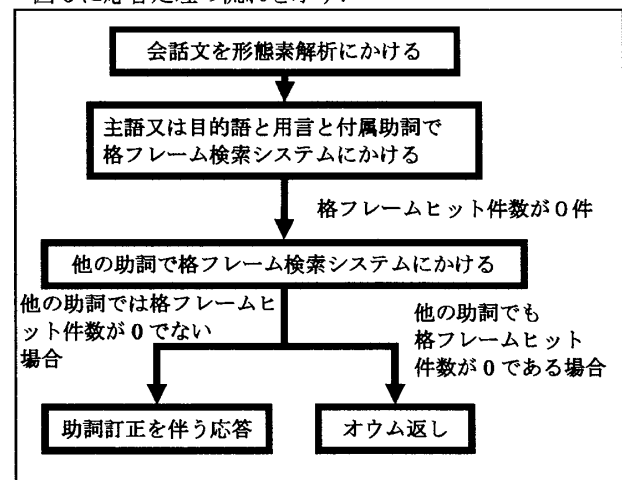


図 6 応答処理の流れ

まず、会話文に対し形態素解析を行い、名詞と動詞、付属する助詞で格フレーム検索システムにかける。例えば「図書館は出発する」という文では、「図書館」と「出発する」に分け、「は」で格フレーム検索システムにかける。次に、格フレームヒット件数が 0 件であった場合には、他の助詞で再び格フレームにかけ、最も格フレームヒット件数が高い助詞で助詞訂正を伴う応答を返す。ここで、他の助詞とは代表的によく用いられる「が」「を」「で」「に」「と」「の」「へ」「まで」「から」「より」とする。

例文の場合、他の助詞で格フレーム検索システムにかけ

た結果、助詞「に」を持つ場合が最も格フレームヒット件数が高くなるので、『「図書館に発着する」ではないですか?』と応答を返す。

また、助詞の誤りではなく、主語や目的語の関係が非常識な文に対しては、他の助詞で格フレーム検索システムにかけても非常識だと判断されるので、オウム返しで会話文を聞きなおす応答処理ができるように対応している。例えば、「マフラーを食べる」を応答処理にかけると、「マフラー」、「食べる」、付属助詞「を」、他の助詞「が」等で格フレーム検索システムにかけても格フレームヒット件数が0件で非常識だと判断されるので、『「マフラーを食べる」ですか?』と応答する。

5.2 評価

アンケートから助詞の付与を誤った文を100文用意し、応答文を評価した。

5.3 評価結果と考察

図7に助詞訂正を伴う応答の評価結果を示す。

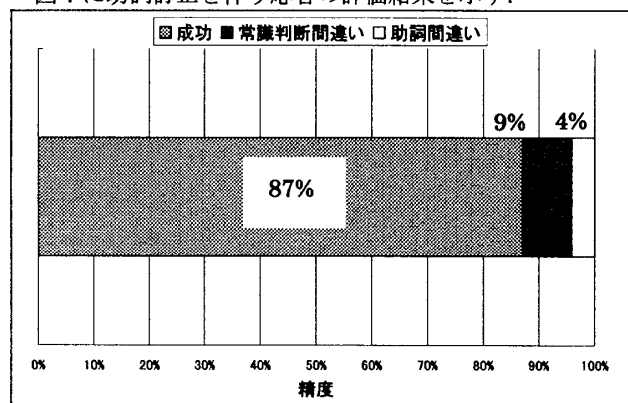


図7 応答文の評価

図7の中の「成功」は、助詞訂正を伴う応答又はオウム返しを行った評価データである。例としては、「耳に澄ます」を『「耳を澄ます」ではないですか?』と応答を返すことができた。

「常識判断間違い」とは、名詞と用言で格フレームにかけた際に格フレームヒット件数が0件ではなく、他の助詞で格フレームにかけなかった間違いである。例として「電話に鳴る」は「電話に音が鳴る」などの文がWeb上に存在し間違いを起こした。その他「温泉が入浴する」、「台風で近づく」などがあった。

「助詞間違い」とは、名詞と用言、付属する助詞で格フレームにかけ格フレームヒット件数が0件で非常識だと判断したため、他の助詞で格フレームにかけたが、助詞の選択が適切ではなかった間違いである。例として、「顔なじみで話す」は「顔なじみに話す」と助詞を変換したかったが、「顔なじみの人が話す」がWeb上に多かったので、「顔なじみの話す」と間違えて変換した。他の例に、「机が使う」、「評価の採取する」などがあった。

今後は、助詞の「の」に変換する場合には、「の」の後に付属する名詞を考慮して、応答処理を行う必要が考えられる。

6. おわりに

本システムにより非常識文の抽出が全体として85%ほど可能になり、システムとして十分有効な精度は示したと考えられる。また助詞訂正を伴う応答も、87%の精度より、

非常識文に対応するシステムを提案でき、その有効性を評価により示したと考える。

しかし、「主語」と「目的語」の関係を考慮した常識文の抽出はできていないため、今後の課題として残る。また、場所名詞を含む文に関しては、Webから自動構築された大規模格フレームのデータベースに少なく、非常識文抽出が十分行えないという問題があったので、場所名詞に関しては別にデータベースを作るなどの方法で解決を図ってきたい。

その他、曖昧文という分類を新たに加えた事により、コンピュータの常識が人間の考える常識に近づいたと思われる。今後は、曖昧文は格フレームヒット件数を用いる方法では網羅しきれないコンピュータの曖昧文を、人間の考える曖昧文に近づけていく方法を考える必要がある。

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った。

参考文献

- [1] 河原大輔, 黒橋禎夫, “高性能計算環境を用いたWebからの大規模格フレーム構築”, 情報処理学会, 自然言語処理研究会, 171-172, pp. 67-73, 2006
- [2] NTTコミュニケーション科学研究所監修, 「日本語語彙体系」, 岩波書店, 東京, 1997
- [3] Google 検索ページ <http://www.google.co.jp/>