特　集「人工知能学会・情報処理学会共同企画―第２部「人工知能における人道とは」―」

対話システムと倫理

Ethical Issues in Dialogue Systems

東中　竜一郎　日本電信電話株式会社 NTT メディアインテリジェンス研究所
Ryuichiro Higashinaka　NTT Media Intelligence Laboratories, NTT Corporation.
higashinaka.ryuichiro@lab.ntt.co.jp

1. 対話システムにおける倫理的観点

今年の 3 月, Microsoft 社のチャットボット「Tay」が，人種差別や陰謀論を含む不適切な発言を行ったとして停止されたことは大きな話題となった．この騒動は，一部のユーザによって，不適切な発言をするように教え込まれたことが原因であるが，考えてみれば，このところ対話システムの発言がネット上でニュースになることが多い．例えば，Siri や「りんな」がこんな発言をした，といったことが話題になったりしている．今はまだ対話システムが日常的に使われているわけではないが，今後，日常的に用いられるに従って，その発言の影響力は大きくなっていくだろう．

対話システムは，ユーザのそばにいて，ユーザに直接的に働きかけることができたり，ユーザの個人的な情報にアクセスすることができる．対話システムの構築者は，その影響力を自覚して，システムの発言が社会（ユーザを含む）にとって良い影響を与えるように努力する必要があるし，対話システムがもし何らかの組織のためにつくられたのだとするならば，その組織に役立つようなものにしていく必要がある．

以下に，対話システムを構築する際に考えておくべき倫理的な観点についていくつか述べる．これらの観点は我々が対話システムを構築するにあたって常々気を付けている点である．なお，これらの観点は全く網羅的ではないことに注意されたい．

1·1　発言内容の適切さ

対話システムの構築者は，システムの発話が社会・組織にとって良い影響を及ぼす発話となるように努力する必要がある．また，システムの発話による悪影響をなるべく軽減するようにすることも必要である．特に，下記の発言は行わないようにする．

誤った情報を含む発言　ユーザはシステムの情報に基づいて行動することがあるので，誤った情報に基づいた発話や真偽が明らかではない発話は行わない．また，音声認識技術や言語処理技術は，精度が 100％にならない技術であることを自覚し，重要な判断はシステムが勝手に行わないようにする．

不要な情報を含む発言　対話システムがユーザとやり取りするにあたっては，社会の利益と組織の利益を考慮する必要があるが，最優先されるべきは社会の利益であろう．よって，組織の利益を追求するあまり，ユーザにとってメリットがない宣伝や情報のフィルタ，価値観の押付けをしない．

不快な表現を含む発言　ユーザを不快にさせるような暴言や誹謗中傷は行わない．ユーザを罵倒したり，乱暴な言い回しをしたりしない．卑猥な表現は用いない．また，人種，国籍，宗教，性別，性的指向，障害などに基づいて個人やグループを中傷するヘイトスピーチを行わない．

反社会的な発言　社会における普遍的な価値観を否定することは，社会にとって悪影響を及ぼす可能性がある．よって，社会的に善とされているものを否定したり，社会的に悪とされているものを肯定するような発話はしない．例えば,「平和は良くないですね」や「地震は嬉しいですね」といった発話はしない．

議論を呼ぶ発言　特定の話題（特に，政治的な話題）について言及すると，議論を呼び，システムの考えが，システムを構築した組織のそれと結び付けられ，組織の利益を損ねる恐れがある．よって，物議を醸し得る話題については，明言を避けるか，組織の考えを正しく伝えるようにする．

「自分が言われて嫌なことは対話システムにも発話させない」というのが大原則である．これは「黄金律テスト」と呼ばれる．チャットボットだからといって，好き勝手に発話してよいわけではない．また，セブンステップガイド [Davis 99] に習い，「システムの発言を第三者の前で擁護できるか」,「システムの発言を同僚が聞いたら何と思うか」などの観点からも確認しておくのがよい．

1·2　プライバシーの保護

対話システムの利点の一つは，文脈（対話履歴）を用いて，短い発話で物事を済ませることができることであり，そのためにはユーザとのやり取りを蓄積し，対話に活用していくことが重要である．しかしながら，ユーザとのやり取りはプライバシーを多く含む．例えば，ユーザはシステムに対して，他人に言えない相談をするかもしれないし，誰かの個人情報や会社の重要情報を言うか

もしれない．プライバシーを適切に保護しないと，ユーザは対話システムを安心して利用することができない．基本的には，ユーザとシステムのやり取りに含まれる情報は個人情報と認識し，事前に提示した目的を逸脱して使用したり，本人の同意なしに第三者に開示・提供しないようにする．加えて，以下の点にも留意する必要がある．

利用目的の説明　対話システムはブラックボックスになりがちなので（対話システムは，何が伝わったのか・何ができるのかがわからないとしばしば批判される），蓄積されたデータがどのように扱われているのかについて，ユーザは不安に感じやすい．システム構築者は，データの利用目的について，十分説明する必要がある．例えば，音声認識率の改善，発話理解の改善（ユーザの身の回りの単語が理解されやすくなる），参照表現の理解（「あれ」や「それ」で通じるようになる），情報推薦精度の改善，ユーザと同調するための言い回しや単語の使い方の学習，などのメリットについてユーザに理解してもらう必要がある．

個人情報に基づく発話　ユーザ以外の第三者がその場にいる状況で，対話システムが勝手にユーザのプライバシーに触れる発話を行う可能性にも注意する．また，一度話したけれど，もう触れられたくない話題というものもあるだろう．ユーザの個人情報について発話しないモードを準備したり，対話履歴をリセットできるようにしたり，保存されたやり取りをできるだけわかりやすく閲覧できるようにし，選択的に対話履歴を削除できるようにするほうが良い．

1·3　関係者の保護

対話システムを世の中で継続的に使ってもらうにあたっては，関係者，特に，ユーザおよびデータ作成者に十分配慮する必要がある．具体的には以下の点に留意する．

人体への影響　対話システムとのやり取りは，見方を変えれば人間を用いた実験である．このことを認識し，ユーザに不要な負荷がかからないようにする．例えば，ユーザを対話データ収集の道具としてむやみに利用したり，ユーザの発話にわざと否定的に答えて反応を試すといったことはしない．

データ作成者の権利　対話システムの構築には一般に大量のデータが必要である．これらのデータはインターネット上のコンテンツをクロールしたり，クラウドソーシングを用いて収集されるが，その際には，コンテンツの著作権を保護することはもちろん，クラウドワーカーに適切な報酬を払うように心掛ける．特に，対話データをクラウドソーシングで収集する場合は，ユーザに発話を考えてもらうなど創作的な要素が強いため，相応の対価を支払う必要がある．

2. より良い対話システムに向けて

情報処理学会誌 [東中 16] でも述べたとおり，対話システムを構築する手法は主に三つある．ルールベース，抽出ベース，そして，生成ベースの手法である．ルールベースの手法でシステムを構築する場合，すべて人手で発話を記述するため，不適切なシステム発話を行わないように制御することは可能である．しかしながら，抽出ベースや生成ベースの手法を用いる場合，適切さを欠いた発話をしてしまうことは起こり得る．

不適切な発話の一部分はブラックリストを用いることで検出できるだろう．しかし，検出が困難な事例も多い．近年，ヘイトスピーチを検出する研究がなされているが，どの手法であってもおおよその精度は F 値で 60 ～ 70% 程度であり [Warner 12]，十分ではない．また，文脈の理解が十分ではないために，システムが不適切な発話を言うように誘導されてしまう可能性もある．例えば，「地震は好きですか」というユーザの発話について，発話がうまく理解できなかった結果，誤ってシステムが「はい」と答えてしまうと，反社会的な発言となってしまう．ユーザが提示する命題が社会的かどうかという判断をしなくてはならないが，このような判断は常識的知識が必要であるため，現状では実現が難しい．

対話履歴の利用については，ユーザが納得するようなメリットを，対話システムの研究者は実現していく必要があるだろう．倫理的な課題は多いが，一つずつクリアしていき，社会に役立つような対話システムを目指したい．

◇ 参 考 文 献 ◇

[Davis 99] Davis, M.: *Ethics and the University,* Psychology Press（1999）

[東中 16] 東中竜一郎：対話システム研究の動向―対話システムは次世代のインタフェースになるか―（特集「人工知能学会共同企画―人工知能とは何か？」），情報処理，Vol. 57, No. 10, pp. 972-973（2016）

[Warner 12] Warner, W. and Hirschberg, J.: Detecting hate speech on the world wide web, *Proc. the Second Workshop on Language in Social Media,* pp. 19-26（2012）

2016 年 8 月 17 日　受理

著 者 紹 介

東中　竜一郎（正会員）

1999 年慶應義塾大学環境情報学部卒業．2001 年同大学院政策・メディア研究科修士課程，2008 年博士課程修了．2001 年日本電信電話株式会社入社．現在，NTT メディアインテリジェンス研究所勤務．しゃべってコンシェルの質問応答機能や雑談対話システムの研究開発に携わる．博士(学術)．情報処理学会，言語処理学会，電気情報通信学会各会員．