

次世代音声自動翻訳への 新たな挑戦

中村 哲さん

(奈良先端科学技術大学院大学・教授)に聞く



今回は、日本の音声自動翻訳研究を第一線で牽引してきた中村哲教授の研究室を訪ねました。人とコンピュータのコミュニケーションを支援する多様な技術の教育・研究に取り組んでおられる奈良先端科学技術大学院大学データ駆動型サイエンス創造センター（DSC）センター長の中村先生にお話を伺いました。

インタビュアー：CIEC 会誌編集長 横川博一（神戸大学）

自動翻訳の夢

横川（編集長） 今日はどうぞよろしくお願い致します。早速ですが、中村哲先生と言えば、音声自動翻訳のご研究を長らくされてきておられますね。私も大学院生の、情報処理学会や言語処理学会などにも出かけていたんですけども、機械翻訳のブームが来ていた時期でした。先生が自動翻訳とか翻訳システムに興味をもたれたきっかけはどのようなことだったのでしょうか。

中村哲先生 どこからお話をすればいいでしょうかね。翻訳に関しては、1986年に当時の郵政省の支援で、ATR 国際電気通信基礎技術研究所で自動翻訳の研究が始まったのがきっかけです。私は当時はシャープの研究所にいたんですけども、そこからの出向という形で自動翻訳の研究に携わるようになったんです。当時は音声認識の研究をしていました。大学では画像系の研究をしていたんですが、卒業して会社で研究開発をしていくのにどういうテーマがいいのかなと当時いろいろと考えていて。音声理解、つまり話したことを理解するっていうことがいったいどういうことなのか、そういうことが使えるよう技術を作るということに興味を持ったんです。いろいろと調べたんですけども、当時シャープの研究所ではワープロっていうのを作っていましたが、声での入力を可能にする音声入力ワープロの研究を通産省（当

時）のプロジェクトとして行っていたんですね。そこに希望して入れていただきました。そうしているうちに、1896年にATRが発足し自動翻訳電話を研究開発する研究所に出向することになり、音声認識に携わることになりました。そこがきっかけなんですね。実際には翻訳の研究を当時はしていませんでした。3年で出向が終わりシャープに戻って、1994年に奈良先端大に移ったあと2000年にATRに戻る誘いがあり、それでATRに戻りました。音声処理研究の室長をしたあと、2005年から自動音声翻訳の研究所の所長として全体の研究を担当することになりました。私個人としては先程言ったように音声理解の研究をしたかったので、自然言語の勉強や研究もフォローしていたのですが、翻訳の分野も面倒をみるようになってそこから、自動音声翻訳や通訳とはなにかを深く考えるようになりました。

横川 音声認識に対する関心がATRでの自動翻訳プロジェクトの中で化学反応を起こして、日本の自動翻訳技術の発展を牽引されてきたということがよくわかりました。しかし、私たちは「翻訳」って簡単に言いますが、相当な苦労があったのではないかと思います。

中村 そうですね。機械翻訳というのはテキストをテキストに変換する技術です。書かれたものは誰が書いても

同じテキストなんですけども、音声言語の場合、声は話すたびに微妙に違い、コンディションによっても違います。また、人や地方によっても違います。そのあたりのことを吸収して、文字化していくというのが音声認識の難しいところです。

横川 音声認識というのは、最初に商品化された時にはその人の声を記憶させて認識するようなシステムでしたよね。

中村 昔はそうでした。話者を決めてその人の声だけで動くような音声認識だったんです。しかし、2000 年ごろからは、誰の声でも認識できるようになってきた。2005 年ぐらいには誰の声でも音声認識がかなり動くようになってきました。機械翻訳と音声合成を上手く統合し、いろいろな実データを集め性能改善を行ったところから、段々音声翻訳っていうものがリアリティを持ってきました。

自動翻訳を可能にした新技術

横川 それは社会的な高まりといたしますか、ニーズがあったからでしょうか。

中村 ひとつはハード面の進化です。大規模データと機械学習などの技術により性能が上がってきて、小型化もできるようになってきました。もうひとつは、ネットワークの進化です。携帯電話が進歩して、それを端末にしてサーバーリモートにおいて、あるサーバで複雑な計算をするということが実際にできるようになったんですね。

横川 それが商品化されて普及していったということでしょうか。

中村 2007 年には携帯電話でそのサービスを入れて、軽い前処理をしたあとそのデータをサーバに送って、認識翻訳あるいはその音声合成の処理をしてまた端末に返すといったようなサービスができるようになったんです。そこからかなり実際に商品化・商用サービスが始まったんですけど、かなりそこでリアリティができて実際に使ってもらう方々が急激に増えていきました。

横川 それは今も発展し続けているわけですね。

中村 2010 年ぐらいからですかね、深層学習の技術が



発展しましたので、それによって各モジュールの性能が飛躍的に向上して、今までよりもさらに性能がよくなりました。

横川 なるほど。私が大学院生だった 90 年代くらいには「事例ベースの翻訳」というような考え方が出てきましたね。それまでは文法規則を 100 も 200 も機械に入れてなんとか文を解析しようというようなやりかたのパラダイムが変わりましたね。

中村 そうですね。ルールやパターンをベースに翻訳する技術ですね。当時、京都大学の長尾真先生がまず提唱されたのが、用例を使って翻訳する方法です。ルールではなくて、たとえば英語のこのパターンが日本語のこのパターンに対応するというのが見つければ、それを代入して翻訳ができるというものです。こういったことをベースにパターンを統計的に学習する技術ができて、2000 年代の主流の技術になりました。これがフレーズベース統計翻訳という技術です。その技術ができたおかげで、対訳ペアを大量に集めればそこから翻訳ができるようになったわけです。

横川 同じようなことが、人間の言語処理でも統計的学習の役割の重要性が言われだしていますが、やはり機械も人間と同じような性質をもつことができれば翻訳精度も上がるということですね。

中村 そうですね。どれぐらいの頻度でこのフレーズのペアが対応しているかっていうのが分かればその確率を求め、あとは翻訳語の文の単語列の自然性の確率を統合することで翻訳することができるようになりました。それが 2000 年代の技術で、我々はいち早くその方向に切り替えて翻訳の性能を高めることができました。

横川 それが今はもっと進歩してきているわけですね。

中村 今はその延長上ですが、統計ベースではなく、深層学習ベースで高性能な翻訳ができるようになってきました。性能はかなり上がってきています。ルールベースや用例のルールベースがあって、そこからフレーズベース統計翻訳になり、そして深層学習に基づく機械翻訳へと進んできています。

ニューラル機械翻訳って？

横川 中村先生のホームページを拝見しますと、音声同時通訳、ニューラル機械翻訳というキーワードが出てきますね。それに、アフェクティブとか感情とか、パラ言語処理といったキーワードも。ニューラルってというのが機械翻訳とどう結びつくのか興味深く感じたんです。

中村 まず、音声翻訳について話します。音声から音声に翻訳するってということがどういうことなのかを考えてみましょう。書き言葉で文字になったものを文字に翻訳するのに比べて、話し言葉の音声を通訳することは非常に難しいんです。どういうものかって言うと、今は音声認識して翻訳して音声合成するっていうような形のものが多くはありますが、話し言葉と書き言葉には違いがあって、話し言葉にはその強調とかイントネーションとか、あるいは感情とか。それからその人の声の個性とか、そういうものがあって、そういうものが相手のその翻訳された声にそのまま対象言語に乗っていくっていうのが、一番自然なコミュニケーションですよ。怒ったら怒ったままの自分の声で相手に伝える。そういうことをやるっていうのが speech to speech のゴールになるんですね。だから、そういった感情を抽出して、その感情を翻訳語の音声に乗せていくとか、あるいは、“あなた”と“わたし”などで、“あなた”の部分に強調があったら、その“あなた”の“You”の部分に強調するとかですね、そういったことが音声から音声の翻訳の場合にはとても重要になるんですね。ですから、そういう意味で、感情とか強調とかといった情報をどう相手の言語で表現していくのか、どう変換していくのかというのがすごく重要になると考えています。そこが文字から文字の翻訳にはない部分になると思うんですね。

横川 興味深いですが、それってすごくチャレンジなことですね？

中村 そうですね。まあ今ちょっと僕らが音声で強調さ

れている部分を見つけて、相手の言語、つまり対象言語で強調する部分を、決めていくわけですが、1単語が1単語に対応するとはかぎらないんですよ。ここら辺もすごく面白くて、すごく重要な技術です。また、原言語で強調あるいは感情が付加されても、対象言語では、それを音声で強調しなくても文字で強調してもいいわけです。だから表現をそのままじゃなくて、いろんな形で表現し得るので、それをどういう風に組み合わせを選んでいくのかっていう問題もあります。今、そういうことも研究しています。文字と音声のパラ言語との関係を、全体として人間が伝えたい情報として捉えて、それを相手言語でどう組み合わせで表現するのかっていうのがすごく重要でもしろい課題だと思っています。

横川 実際のコミュニケーションでは、いわゆる字面だけではない部分の役割が大きいですね。ほかにもチャレンジングな面はあるのでしょうか？

中村 深層学習です。つまり、ニューラルネットワークです。ニューラルネットワークっていうのは基本的には、現在は多層型のニューラルネットワークを指していて、層が“入力に入れる層”とそれから複数の中間的な層があって、それをまとめる出力層から出力が出てくるんですが、その中間的な層に非線形な関数が埋め込まれています。入力を入れて与えられた出力が出るようにニューラルネットワークのパラメーターを学習するというような構成になっています。1987年にATRにいた際に、ニューラルネットワークの第2次ブームが来て、私もその研究をしていたのですが、当時の計算機ではなかなかうまく学習することができなかったり、その学習するためのデータが足りなかったりしました。

横川 その後、計算機は発展したわけですね。



中村 2010 年以降になると計算機も速くなり、実際に性能がすごく改善しました。ニューラルネットワークの中心人物はジェフリー・ヒントンという先生なんです。データ量と理論と計算機的能力がうまくフィットして、音声認識、機械翻訳、画像処理の分野で画期的な性能改善をもたらしました。その技術をさらに発展させていろいろな研究を進めています。

同時通訳者の行動に学ぶ

横川 なるほど。どう関連しているかがだんだんわかってきました。それに加えて、同意通訳となると、さぞ大変なんだろうなと感じ始めました。

中村 同時通訳ですが、これまで音声翻訳っていうと、一つの文章、一つの文、あるいは一つの発話を入力してそれを音声認識して翻訳して合成して出すという形でした。つまり一つの発話を文末まで待ってから翻訳して音声合成するという処理でした。ですから、講演のように何秒も何十秒も続くような発話については翻訳出力が出てくるまですごく時間がかかります。当然のことながらすばやく出力するようにしたいですね。また、目標としては、人間の通訳者のような高度な処理を目指したい。どんな処理をしているのか、計算機でどのぐらい真似ができるのか、人間の通訳者には記憶の限界があるので機械が上手に支援できないか、あるいは、人間を超える同時通訳ができないか、そういうことを目指して同時通訳の研究を科研費などをいただいて進めています。

横川 具体的にはどんなふうに研究を進めておられるのでしょうか。

横川 TED Talk というプレゼンテーションがありますね。英語でも日本語でも行われていますが、同じプレゼンテーションに対していろんなレベルの通訳の方にしてもらって、そのデータを比較したり、データ使って遅れの少ない通訳の方法の研究、深層学習によるシステムを作ろうと考えています。通訳データ構築の目標としては 400 時間分くらい集めたいと思っています。

横川 大量のデータから何らかの要素を抽出するっていうことになるかと思いますが、なにか「あたり」をつけて要素を抽出していくのでしょうか。

中村 そうですね。たとえば日本語から英語の通訳の場合は、日本語の動詞が一番後にきますよね。英語は後で

すね。そうすると日英翻訳の場合は、動詞が現れる文末近くまで待たなければ翻訳できないということになるんですが、人間の通訳者は途中である程度予測 (anticipation) して訳したり、長い文章だと思うと 1 回文を止めたりするんですね。そういった機能を機械で実現できないかということをやっています。また、その良し悪しは別として、通訳者は重要なことだけ訳して、全部は訳さないですね。これは興味深い行動で、機械翻訳でもそういう不要な部分を落としていく、そういうこともやるということもやってみたい。

横川 人間の通訳者の行動というのもおもしろいですね。

中村 こんなこともあります。通訳者が実際に通訳をしている中で、通訳者を支援するようなシステムっていうのもあるんです。実際には通訳者は 2 名セットでやられていて、一人は休憩している間にも聞いていて、たとえば数字とか重要なワードをメモに残しています。そういったことも自動でできるようにするとおもしろいのではないかと考えています。

横川 先生のお話を伺っているとワクワクします。機械で同時通訳がほぼ可能な時代にもなっていると言っているのでしょうか。

中村 まだだだと思いますね。同時に、なるべく遅れなく翻訳するっていう研究は難しい研究課題であることは確かです。ただ、世界の何か所かで先進的な研究がなされています。そこではより性能の高い新しいアルゴリズムを生み出すために激しい競争が繰り広げられている状況です。

機械翻訳のリアリティ

横川 1980 年代にも、そういった具合に機械が翻訳してくれるといいなあと思っていました。当時はどういったことを目指して機械翻訳のシステム開発に当たられていたのでしょうか。最近なら東京オリンピックに向けて自動翻訳研究が盛り上がりを見せているように思いますが (笑)。

中村 当時 (2000 年代) は、まだ、当然オリンピックも決まっていませんでした (笑)。私たちの ATR のプロジェクトは 2008 年から NICT (情報通信研究機構) のプロジェクトとして引き継がれました。当初から、実際

なるべくたくさんの人たちに使ってもらえる音声翻訳装置システムを作ろうということで、旅行での利用を目標としました。日本人が海外に行く場合と海外の人が日本に来る、そういう場合に使われるような音声翻訳装置です。そういった時に出てくる発話っていうのは比較的短い発話ですし、音声認識もしやすいし、機械翻訳もしやすい。本語では You とか He や She といった主語が抜けますが、旅行という場面では基本的に 1 対 1 で発話が行われますので、だいたい You でいけるわけです。

横川 なるほど。そうしますと、旅行会話の音声翻訳、言うなれば旅行会話を中心に音声翻訳システムの開発を進められたんですね。それはその後どのように展開されていったのでしょうか。

中村 2010 年には iPhone や Android でそういう基本的な音声翻訳のトライアルサービスを無料で公開してデータを集めさせてもらいました。そして今、オリンピックが決まりましたから、そういう技術を実際に使ってもらえるチャンスが実際に来たわけです。ただ、私は 2011 年から大学に戻りましたので、この仕事は NICT にいる仲間が頑張って進めてくれています。

横川 海外でももちろんこのような動きはあるのでしょうか。

中村 海外でも、外国語の翻訳という意味では非常に盛り上がってしまっていて、日本では我々の技術をベースに 2007 年に携帯電話の通信ネットワーク音声翻訳の商用サービスが始まりました。その後、2008 年でしたか、Google は音声入出力の Google Translation をリリースしました。実は日本の方が早いんです。現在では、マイクロソフトや中国の企業がサービスを始めています。海外の人が日本で、音声翻訳サービスをよく使っているのを見かけます。

横川 先ほど先生が言われておられた、短い文章で、同時に機械翻訳・音声翻訳をするという課題はどう進んでいくことになるのでしょうか。

中村 短い文章で翻訳するというのは、現状すでになんか確立された技術で、あとは語彙などのデータを集めることで課題はかなり解消されると思っています。ですが、同時通訳、これは私たちの最終的なゴールだと思っているんですが、同時通訳のような状況の技術っていう



のはまだまだ足りません。実際の通訳者は、翻訳する対象の発話を理解しながら、発話者はまだ喋っていないのに通訳者がその先を翻訳しているといったようなことが結構頻繁に起こっていますが、やはり中身を理解してこそその翻訳であるという点も見逃せません。

横川 人間は、これまでの経験にもとづいて、常に先予測しながら相手の発話を処理していますね。それも効率性を図る意味では、重要な性質ですね。

中村 そうですね。そういう状況までもっていくとなると、内容の理解とか背景知識など、発話者が喋っていること以外の情報も使いながら通訳していかなければいけませんから、そこまでやろうとすると、まだまだやらないといけないことがたくさんあります。今、そういう研究を頑張って進めています。実際にいつできるのかを言い当てるのは難しいですが、これから、今まで難しかったようなことがいろいろとできるようになったりしていくと思います。大量のデータが必要になりますが、いま収集しているデータを利用していろいろとおもしろいことができると思います。そういう時代になってきたと思います。

英語教育は要らない?!

横川 ちょっとお決まりの質問になって恐縮ですが、ひとつ先生のお考えをお聞かせください。私は外国語教育を専門としていますので、新聞の取材も受けたこともあるんですが、こういう翻訳システムが進歩してきて、外国語教育・外国語学習が要らなくなるんじゃないかと。こういう問いに対しては、先生のようなお立場からは、どのようにお考えでしょうか。

中村 たしかによく尋ねられます (笑)。音声翻訳システムがあったらもう英語は勉強しなくてもいいんですかって。学会などでいろいろな国に行きますが、自分が言

いたいことをきちんと伝えるためには英語なりその言語をきちんと勉強して、face-to-face でリアルタイムで伝えるようになってもらいたいと思います。そして、音声翻訳技術は、英語を勉強するとき、英語以外の中国語やドイツ語とかフランス語など第二、第三の外国語を勉強するのがちょっと大変だなというところで利用する、あるいは、内容の難しい講演や会議の同時通訳として利用するのがよいと思います。心を割って話すということは機械ではやはり難しいですから、自分でその外国の人と本当に親しくなろうなどと思うときには、やはり自分で話すことが重要です。少なくとも現状ではこのようなことが言えると思いますし、しばらく先もそうだろうと思います。ですから、こういう技術を使いながらその外国語を勉強していくというのが私は理想だと思っています。

横川 私も同じ考えです。こういう技術は増々発展した方がよいと思っています。学習ツールとして使えますから。みんな、英語が嫌いだと言っている人も含めて、密かに力を付けたいと思っていますからね。こういう技術が発達してくると、夜中でも早朝でも（笑）、自分で勉強して、今度実際に自分で喋ってみようというモチベーションにつながっていくといいなと思いますね。

コミュニケーションを支える技術

横川 これまで先生のお話を伺っておりますと、事例ベースの翻訳あたりから、かなり人間がやっていることに近い形で翻訳システムを作っているように思います。ですから、話者の感情を理解したり、意図を理解したりということがとても重要になってきますね。翻訳システムも、もはや単なるツールではなく、先生の研究室のキーワードのトップに挙がっている「コミュニケーション」そのものと言えるように思います。先生の研究室では、助教の田中宏季先生らと一緒に、コミュニケーション支援の研究にも取り組んでおられるようですが、そのあたりのこともご紹介いただけないでしょうか。

中村 対話の音声翻訳は翻訳機を通して異なる言語を話す人々のコミュニケーションを実現する技術です。一方、最近では、機械が人間の代わりに応答する対話システムの技術の研究も進んできました。人間が何か問いかけをした時に適切な応答をするような対話システム、これは人工知能そのものと言えます。具体的には、情報提供をする対話や、高齢者との傾聴対話を目指したシステ



ムの開発が進んでいます。現在は、テキストレベルで対話応答するものが多いですが、今度は、人間が音声で表出した強調とか感情、そして言葉の内容を理解することができるようになり、対話システムの方は究極の人工知能のような技術になっていくと考えています。

田中先生と一緒に進めているのは、自閉症の子供達にコミュニケーションの仕方を教える人間のカウンセラーの作業を、対話技術を使って行う技術です。これをソーシャルスキルトレーニングといいます。自閉症の方たちは、比較的機械とコミュニケーションする方が得意だということもあって、比較的抵抗なく使ってもらえます。普段なかなか SST のトレーニングを受けられないけども、家でゲーム的にやることでコミュニケーションスキルを上げるという、そういう研究ができるんじゃないかと思っています。基本的なコミュニケーションとか人間の言葉や非言語の情報を扱う技術、それからそれらを使って自閉症の方々のトレーニングに役立つ技術を作るといことに取り組んでいます。

横川 教育にも、たぶんに応用できることがありそうですね。

中村 はい、先程の英語教育の話もありますし、今の自閉症の方々への教育とか、健常者でもコミュニケーションが苦手な大人もたくさんいますので、そういうところに、英語も第二言語も含めて教育の場面に私たちの研究は使ってもらえると思っています。

横川 この学会の中には教育に携わっている小学校の先生もおられたり、中学校の先生もおられたりするんですけど、最後に、そういう方々にメッセージをいただけないでしょうか。

中村 まだ完全ではないんですけども、人間が言ったことを聞きとってそれに対して適切なアクションをする

ような技術の研究が進んでいます。今後、教育への応用もある程度できるようになってくると思います。これからやはり子供たちそれぞれに合わせた課題提示、子供達が回答した内容の自動的な理解、それに応じた回答、また、成長に合わせた教育などができるとよいと思います。また、ゲーミフィケーションと言って教育の中にゲーム性をいれて効果を高める要素なども、宿題の中にとり入れて、楽しみながら勉強ができるようになると思います。そして、このような技術により、教育に携わっておられる先生を多忙なルーティンワークから解

放し、より本質的な教育に携わる時間が増えると素晴らしいと思います。

横川 今までの学習ソフトとは全然違う、もっと対話的に、ゲーム的に学べるような時代がもうそこまで来ているということですね。子供達も先生方ももっと楽しい教育現場を創ることができたらいいですね。今日はお忙しい中、お時間を割いてくださり、ありがとうございました。