機械がことばを理解し、話すことについて

土屋, 智行
九州大学大学院言語文化研究院 : 助教

https://doi.org/10.15017/4481576

出版情報：言語文化叢書. 23, pp.3-17, 2020-11-13. Faculty of Languages and Cultures, Kyushu University
バージョン：
権利関係：

目 次

＊土屋助教は 2016 年度から 2018 年度まで 3 年間本講義を担当したが，ひとつの原稿に内容を集約させているため，担当初年度である 2016 年度のところに掲載する．劉准教授は 2017 年から 2018 年度まで 2 年間本講義を担当したが，2 年とも内容が同一であったため，同じく担当初年度である 2017 年度のところに掲載する．

機械がことばを理解し、話すことについて

土屋 智行

前置き：「ことばの科学」と「言語学概論」の違い

授業に入る前に、「ことばの科学」という、この科目の名前について私なりに考えてみたいと思います。

本来、ことばに関する科学的なアプローチには「言語学」という名称が付けられています。したがって、この科目は「言語学概論」にあたるのではないかと私は当初考えましたが、それならわざわざ「ことばの科学」ではなく「言語学概論」という科目名にすることもできたはずです。現在の科目名になったのは「この科目が九州大学基幹教育のフロンティア科目であるから」という理由や「『言語学概論』という名前の授業がすでにあったから」という理由もあるでしょう。もしかしたら学生の皆さんが親しみやすく感じるために「ことばの科学」（「言葉」ではなく、「ことば」）としただけかも知れません。しかし、ここではこの科目名の学術的な立ち位置をしっかり考えてみたいと思います。

まずは「ことばの科学」のシラバス（2018 年度）を確認してみましょう。そこには「異なる言語の中に存在する「法則」「規則」を発見・観察する楽しさ、興味深さを、受講生に体験してもらいます。」と書いてあります。つまり日本語と英語、フランス語とドイツ語とスペイン語のように、2 つ以上の言語を比較したときに、それらが共通して持つ（音韻から談話に至る）言語的な規則にどのようなものがあるのかを探し、それを観察する楽しさを学ぶということがこの科目の目的です。人が話すことばに科学的なアプローチが試みられてから、様々な知見が発見され、教育や情報処理に応用されてきました。

これまで学術的な系譜の一端を紡いできたことばへの科学的アプローチ、すなわち言語学ですが、今後はどのような道を辿っていくのでしょう。現代は人工知能や機械学習が学術的な流行となっており、言語学の領域においても、この潮流に強く影響を受けていると言えます。なぜなら、機械がことばを理解し、話すことは、人間が機械とコミュニケーションをとり、機械から情報を得て、機械に仕事をさせるにあたって必要不可欠な機能だからです。現在、機械学習の技術は飛躍的に向上しましたが、未だ解決に至っていない課題も多くあります。これらの課題のどのような点に、言語学はどれだけ関わることができるのでしょうか。

私の担当では、(1) 機械翻訳の技術的発展の概要と課題、(2) 機械翻訳の発展の基盤となる「コーパス」、(3) 人間によることばは機械によることばとどれだけ違うのか、の 3 点についてお話したいと思います。また最後は、これまでのレポートの課題内容について簡単にお話したいと思います。

言語学という領域の中でも、私が専門としている（ゆえに今回の中心的な話題となる）コーパス言語学や自然言語処理は比較的歴史が浅いものであり、一部は言語学のスコープから外れる部分もあるかと思います。しかし、そこからことばに関する新しい科学的アプローチの展開を探る方法も可能なはずです。皆さんにはこの授業の話を通して、ことばを理解し話す機械と人間の関係について、考えを深めてもらいたいと思います。

ことばをあつかう機械の発展

近年の「人工知能」ブームと機械翻訳からみる「ことば」

2018 年の 5 月に、Google が Google Duplex という人工知能アシスタントのデモを公開しました（Google Developers 2018）。デモでは Google Duplex が人間の代わりにヘアサロンやレストランに電話で予約をおこなう様子が音声として流れ、機械であるはずのアシスタントの声や会話の内容がきわめて自然であることに、観衆は驚きを隠せませんでした。実際のところ、全てのプロセスが機械によっておこなわれている訳ではなく、一部は人間による操作が関わっているということが明らかになっていますが、それを差し引いても、昨今の人工知能や機械学習の技術は目覚ましい発展を遂げていることは否定できないでしょう。

もう一つ、ことばに関わる機械の発展の例として、機械翻訳を挙げたいと思います。先ほどと同じ Google が 2016 年にニューラル機械翻訳（Google Neural Machine Translation: Google NMT）をインターネット上に公開しました（Wu et al. 2016）。2020 年現在においても、Google NMT はインターネット上で無料で利用できる機械翻訳システムとして高い翻訳精度と実用性を誇っています。ここで下の日本語文を GoogleNMT で英語に変換してみると（2020 年 6 月時点）、使用される語彙や文法構造は、人間が翻訳したかのような自然さを感じることができます。

Google NMT は、複雑な構造の文にたいして、自然な翻訳にかなり近い文が出力される。

Google NMT outputs sentences that are very close to natural translations for sentences with complicated structures.

2000 年代初頭の機械翻訳ソフトでこの文がどのような日本語文へと翻訳されるのか、あいにく確認することはできませんが、当時と比較しても翻訳の精度は高くなったように感じます。私は父の仕事の手伝いで日本語の論文の一部を、機械翻訳の内容を参考に英訳したことがあります。父が最初に私に持ってきた日→英の機械翻訳の文章は（元の日本語の文章が機械翻訳するには適切ではなかったというのもありますが）自然な英文からは程遠く、結局私が最初から書き直したことを覚えています。また昔の機械翻訳に関するものに、原 倫太郎・原 游による「背面ストライプの浦島太郎」があります。この本は昔話の「浦島太郎」の文章を 15 の機械翻訳ソフトを使って英訳した後、その英訳を同じソフトを使って再度和訳し、各ソフトが出力した文章を切り貼りしイラストを付した本です。この本は不自然な翻訳部分だけを拾い上げて物語をおもしろおかしく編み上げたものなので、本に使われなかった（自然な）翻訳も少なからずあったかと思いますが、それでも当時の機械翻訳の問題点を理解するには十分な情報を提供してくれるものだと思います。昔の機械翻訳システムは、人間による大幅な修正を必要とするものだったのです。

機械翻訳の発展と近年の流れ

2000 年代初頭と比較すると、機械があつかうことばの精度は飛躍的に向上しました。機械翻訳はその典型的なものと言えるでしょう。ここでは、機械翻訳について簡単に述べたいと思います。

機械翻訳とは、コンピュータを用いて元言語の文章の入力を目的言語の文章へと出力する技術を指します。機械翻訳で日本語文を英文へと翻訳するとき、元言語が日本語、目的言語は英語となります。機械翻訳の研究と開発は一定の歴史があり、これまで多くの機械翻訳の方式が提案・実用化され、1980年代から製品が販売されています。過去の機械翻訳は、単語の辞書の作成や文法ルールの構築等に多くの労力を必要としていましたが、近年ではできるだけ人手を介さず、翻訳までの多くのプロセスを機械に任せる方法が主流となってきました。この授業でこれまでに開発されてきた方式を全て紹介することはできませんが、一番最近の方式である統計的機械翻訳とニューラル機械翻訳について、少しだけ説明したいと思います。

統計的機械翻訳（statistical machine translation）は元言語の文章と、その翻訳文がそれぞれ対になっているデータ（対訳データ）を大量に収集し、翻訳モデルと言語モデルという2つの統計モデルに基づいて、入力文に最も近い目的言語の文を出力する機械翻訳の手法です。翻訳モデルとは、元言語の文章を構成する一つひとつの要素（単語、句、等）が、それぞれ目的言語のどの要素に、どれだけの可能性で対応しうるかを算出するモデルです。もう一つの言語モデルは、目的言語に出力される文がどれだけ文として自然・適切かを算出するモデルです。統計的機械翻訳では、翻訳モデルによって入力文の内容に適した出力文の構成要素を探し出し、言語モデルによって出力文の構成要素を適切に変換・調整するというプロセスが実行されます。機械翻訳における入力文と出力文の対応が、対訳データにおける元言語の文章と目的言語の文章の対応に一番近づくように「翻訳モデルと言語モデルをどのように構築するか」が統計的機械翻訳の研究における重要な課題となります。

統計的機械翻訳は、元言語の入力から目的言語の出力までの間に、対訳データを集めたり、入力文を小さい単位に切り分けたり、出力文の自然さを推定して調整する等、複数のステップが存在し、それぞれのステップの構成と組み合わせを人間が考える必要があります。それに対してニューラル機械翻訳（neural machine translation）は、多くの対訳データを必要とする点は統計的機械翻訳と共通しますが、これまで必要だったステップをニューラルネットワークのみで実現しようとするものです。ニューラルネットワークとは、人間や生物の脳神経系を模した情報処理の手法であり、情報の受け渡しをおこなうノードや、そのノード間のつながり方やつながりの強さ、ノードの数や階層の数が定められた構造を持ち、その構造の中で情報を分散並列的に処理する機構です。ニューラル機械翻訳では、まず機械が読み取れるように対訳データをすべてベクトル化し、そのベクトル化したデータをニューラルネットワークに読み込ませ、学習させます。データを学習した機械に入力文を与えると、入力文はベクトルに変換され、その値に最も近い目的言語のベクトルが探索、出力されます。したがってニューラル機械翻訳は、統計モデルを構築する統計的機械翻訳とは違い、「対訳データや入力文をどのようにベクトル化し、どのように読み込むか」というニューラルネットワークの適切なデザインが重要な課題となります。

コーパス：機械翻訳を支える言語データ

近年の機械翻訳の主流である統計的機械翻訳とニューラル機械翻訳は、出力文の精度や問題点等、様々な点で違いがありますが、両者は「大規模な対訳データを必要とする」点で共通してい

ます[1]。統計的機械翻訳とニューラル機械翻訳が現在の機械翻訳の主流となってきた経緯には、機械そのものの性能の向上や、画期的なアルゴリズムの開発が関わっていますが、実は大量の言語データの取得が可能となったことも大きく関わっています。

私達の実生活はたくさんの言語活動に溢れており、私達は様々な場のことばに触れて生活しています。新聞、書籍、インターネットのニュース記事、SNS、日常会話、メール、電話でのやりとり等、ことばに触れない日は無いと言っても良いかも知れません。機械がことばを使えるようになるには、まずこれらの言語活動の場から目的に合った言語データを取得しなければなりません。特定の言語活動の場における言語を収集・整理し、分析に用いれるように構造化したデータをコーパスと言います。機械は大規模なコーパスを使ってことばを学習しているのです。それでは、このコーパスは現在、どのような種類が、どれだけの規模で存在するのでしょうか。また日常に溢れている言語活動は、どのようなプロセスを経てコーパスとなるのでしょうか。ここでは現在のコーパスの現状・構造・構築方法について説明していきます。

コーパスの現状

現在、多くの研究機関や研究者によって多種多様なコーパスが構築されており、それぞれのコーパスは特定の目的に沿って構築されています。一般的な傾向として、コーパスのデータサイズは年々大きくなり、文字だけでなく音声や映像などの情報が加わったコーパスも登場しています。規模に関して言うと、1994 年に公開された British National Corpus (BNC) は約 1 億語でしたが、2008 年に 約 28 億語からなる enTenTen08 コーパスが、2015 年には 150 億語からなる enTenTen15 コーパスが公開されています。日本語でも 2008 年に 4 億語からなる JpWaC コーパスが、2013 年には 100 億語からなる JpTenTen コーパスが公開されています。このような規模の拡大の背景には、インターネットの爆発的な普及と、そのインターネット上から言語データを効率的に収集するクローリング技術の開発があります。またコーパスの種類に関して言うと、データの蓄積・アーカイブ化の技術の進展によって、文字以外の情報（音声や映像など）をコーパス化する動きが活発となっています。たとえば、1996年〜2000年代は英語、日本語、中国語等の6つの言語で、自宅への電話会話を収録し書き起こした CallHome コーパスが公開されました。また国立国語研究所は 2004 年に日本語の話し言葉の音声と詳細な書き起こしデータからなる「日本語話し言葉コーパス」を、2018年には日本語での日常会話の映像・音声・書き起こしデータを収録した「日本語日常会話コーパス（モニター版）」を公開しました。これらのコーパスはいずれも研究のために使われますが、先ほどもお話したように、大規模なコーパスは機械によることばの学習用データとして使われたり、音声・映像を含むコーパスは人と人のコミュニケーションや相互行為の領域での研究にも使われます。

コーパスの構造と種類

コーパスはことばに関するデータを集めたものですが、ただ単純に文字や映像を記録しただけのものではなく、様々な情報やその対応関係を参照できるように構造化されています。一般的にコーパス内の言語データは一定の単位に従って区切られ、その単位ごとにアノテーションが加え

[1] もちろん、それ以前の機械翻訳の方法でも対訳データは重要でしたが、それぞれの言語の専門家が分析した語彙や文法のルールに基づいて機械翻訳のシステムを構築するのが一般的でした。

られています。また、各データにはメタデータとしてレジスターという情報が加えられています。以下にいくつかコーパスのタイプを紹介しつつ、それぞれの用語を説明していきます。

コーパス構築のために収集される言語データには様々なものがありますが、いったいどのようなデータをどれだけ集めたら良いのでしょうか？テレビ、ラジオからもことばは絶えず流れてきますし、新聞、本、雑誌、チラシ、仕事の書類等の紙媒体のものも日常的に目にします。インターネットにおいてはニュース記事やブログ、掲示板、動画サイト、SNS 等でことばを見聞きします。また、テレビであってもニュースとバラエティではことば遣いが大きく違うように、同じ媒体が同じことばを使っているとも限りません。音声を例に取っても、講演等のひとり語り、ニュースの読み上げ、街角のインタビュー、日常的な雑談等では幅がありますし、それぞれの語り口や使っている語彙や言い回しは（程度の差はあれ）異なります。また話者の年齢や性別、社会的属性によってもことばの使い方は違ってきます。一般的に、ことばの研究で言語データを集める際には、多種多様なことばから研究対象を絞ったり、比較をおこないます。このようなメディアの媒体や場面、話者の特性、等のように、ことばを発した物／者や発せられた時代等の特性をレジスターと呼びます。一般的にコーパスは特定のレジスターに絞って言語データが集められており、データそのものにこのレジスターの情報を加えています。

コーパスのレジスター情報は、ことばそのものの情報というより、そのことばの出処に関する情報です。写真にたとえれば、レジスターはいわば写真が撮られた場所やカメラ、時代、撮影者の情報といえるでしょう。このような、ある情報の出処や、情報そのものの特徴に関する情報をメタデータと言います。コーパスにおいて、レジスターは非常に重要なメタデータですが、他にどのようなメタデータがあるでしょう？まず、コーパスの名称は典型的なメタデータの一つですね。他にもコーパスの作成者、使用言語、コーパスの規模（データサイズ、文字・形態素数、付与されている情報の種類、等）が挙げられます。話し言葉であれば、それぞれの発話がおこなわれた時間情報も重要なメタデータとなります。コーパスを使ったことばの研究において、メタデータは研究対象の特定や比較のために頻繁に参照され、使われます。

先ほども述べたように、一般的にコーパスが構築される際には、収集する言語データのレジスターが絞られますが、社会のことばの実態にできるだけ則した内訳となるように、レジスターや規模を調整したコーパスがいくつかあります。たとえば%%%%節で挙げた BNC は、話しことばと書きことばが約 1 対 9 の割合で構築され、更に書きことばも書籍や新聞等の複数のレジスターのデータが収集されています。また、年代も 1985～1993 年に加え、1960～1974 年、1975～1984 年のデータも取り入れられています。このように、多様なレジスターの言語データを、社会におけることばの実態に即して配分・調整したコーパスを均衡コーパスと言います。日本語では「現代日本語書き言葉均衡コーパス」が国立国語研究所によって公開されています。均衡コーパスは、社会におけることばの使用実態の注意深い観察と分析をとおして内訳が検討され、構築されています。社会でことばがどのように使われているか、偏りなく分析・考察するにあたって、均衡コーパスは非常に重要な役割を担っています。

もちろん、均衡コーパス以外のコーパスも、様々な研究において貴重なデータとなります。たとえば通時コーパスは、特定のレジスターの言語データを異なる時代から収集したコーパスで、主に言語変化を研究対象とする歴史言語学の分野で使われます。通時コーパスの構築には、紙媒体の言語資料を電子化して書き起こし、時代によって異なる綴り等の異表記や文法項目を整理する作業を必要とします。今では文字の書き起こし作業に文字認識等の技術が応用されているもの

の、様々な研究のニーズに対応できる規模にするために、専門家の労力を多く必要とするコーパスと言えるでしょう。

また近年大量かつ大規模に構築されている類のコーパスとして、インターネット上の記事やSNS のコミュニケーション等の文字情報を収集・整理して構築されたウェブアーカイブコーパス（Web-archived corpus）があります。%%%%節で述べた EnTenTen15 コーパスや JpTenTen コーパスはその例です。ウェブアーカイブコーパスはクローラーと呼ばれるプログラムを利用してインターネット上の言語データを網羅的に収集しているので、どのような属性を持つ人が発したことばなのか分からないし、校閲等のプロセスを経ていないので誤用や異表記が多いので、均衡コーパスのようにことばの実態を反映したものとは言えませんが、逆に新語や創造的なことばの使い回しを収集できるという特徴があります。また規模がきわめて大きく、機械がことばを習得するための学習用データとして利用されています。

授業の最初で取り上げた機械翻訳システムで使われる対訳データは対訳コーパスもしくはパラレルコーパスと呼ばれます。対訳コーパスは、文学作品やニュース記事等、原文と翻訳文が両方存在する言語データを収集し、構築します。2018年に本学の山村ひろみ先生が中心となって現代ロマンス諸語の対訳コーパスを作成しましたが、これは英文学作品とそのフランス語、イタリア語、スペイン語、ポルトガルポルトガル語、ブラジルポルトガル語、ルーマニア語訳が対応付けられているコーパスで、ロマンス諸語間の様々な文法項目の比較対照研究に有用なものと言えるでしょう。現在は対訳コーパス構築にも、ウェブアーカイブコーパスと同様にクローリング技術によってニュース記事等が利用されることがあります。他のコーパスと大きく違うのは、対訳コーパスは構築にあたって、原文と翻訳を文もしくはそれよりも大きい／小さい範囲で対応させる必要がある点です。

最後に学習者コーパスについても紹介しましょう。学習者コーパスは、日本語の場合は日本語、英語の場合は英語の学習者に特定の課題（会話、レポート、エッセイ、等）を与え、その内容をコーパスとして構築したもので、母国語が学習言語に与える影響や習熟度ごとの語彙・文法の傾向等の研究、より効果の高い外国語教育の検討や検証等に用いられます。学習者コーパスは課題の設定や（母語や習熟度が似通った）協力者の募集、誤用の特定等の作業を必要とし、現在構築されているコーパスの中でも人手の作業の割合が大きいタイプのコーパスと言えるでしょう。

以上のように、研究や用途に応じて様々なコーパスが、多くの研究者達によって開発・利用されています。

コーパス内のことばの情報

コーパスはそれぞれの目的と用途に合わせて独自の言語データの収集・整理・構築がなされますが、ほぼ全てのコーパスに共通するのは、コーパス内のことばが文ないしはそれよりも小さな単位に区切られ、それぞれの単位に特定の情報が付与されている点です。それぞれの単位に付与されている情報をアノテーションと言います。コーパスによって区切られる単位の大きさ、アノテーションの内容や数、方法は大きく違うため、一概に言えない部分もありますが、ここでは自動的なアノテーション付与のツールである形態素解析器と「日本語書きことば均衡コーパス」のアノテーション構造について紹介します。

コーパス毎に言語や用途が異なるので「コーパス内の文をどれくらい小さい単位に区切れば良いか」「どのような種類のアノテーションを付与すれば良いか」について簡単に答えることはで

きませんが、日本語のコーパスの場合、形態素に区切られ、それぞれの形態素に品詞（Parts of Speech, POS)の情報が付与されているのが一般的です。品詞とは、形態素が文の中で担う文法的な役割で、日本語の場合は名詞や動詞、格助詞等のカテゴリーが存在し、全ての形態素はいずれかの品詞に分類されます。形態素解析器とは最も広く利用されている言語処理ツールの一つで、文を形態素に切り分け、それぞれの形態素に品詞情報を自動的に付与する機能を持っています。日本語の形態素解析器にはMeCab, Chasen, Juman, KAKASI 等がありますが、ここではMeCabを例に簡単に説明しましょう。

MeCab は京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究プロジェクトを通じて開発された形態素解析器で、「日本語書き言葉均衡コーパス」やJpTenTen コーパスの構築にも利用されています。MeCab を利用するためには解析用の辞書データベースが必要で、MeCabの場合はUniDicという国立国語研究所が開発したものが一般的に利用されています。

MeCab に文章を入力したらどのような情報が出力されるのか、見てみましょう。「自分の持ち物にはかならず名前を書いてください。」という文をMeCabに入力すると、図1のような情報が出力されます。

```
自分の持ち物にはかならず名前を書いてください。
自分	ジブン	ジブン	自分	名詞-普通名詞-一般
の	ノ	ノ	の	助詞-格助詞
持ち物	モチモノ	モチモノ	持ち物	名詞-普通名詞-一般
に	ニ	ニ	に	助詞-格助詞
は	ワ	ハ	は	助詞-係助詞
かならず	カナラズ	カナラズ	必ず	副詞
名前	ナマエ	ナマエ	名前	名詞-普通名詞-一般
を	オ	ヲ	を	助詞-格助詞
書い	カイ	カク	書く	動詞-一般	五段-カ行	連用形-イ音便
て	テ	テ	て	助詞-接続助詞
ください	クダサイ	クダサル	下さる	動詞-非自立可能	五段-ラ行	命令形
。	。		。	補助記号-句点
EOS
```

図1: MeCabの出力例「自分の持ち物にはかならず名前を書いてください。」

入力した文章が1行目に出力された後、下に解析の結果が表示されています。入力文が形態素に分割され、1 行につき 1 形態素とその情報が出力されます。行の一番左に書字形（形態素の表記）が表示され、そこから右に向かって順番に音形（形態素の読み）、語彙素（形態素から語形変化を取り除いたもの）の音形、語彙素の書字形、品詞が続きます。形態素が語形変化するタイプの品詞である場合、更に活用タイプと活用形が続きます。

このような情報が、一般的なコーパスの内容に付与される基本的なアノテーションとなり、言語学的な分析に利用されたり、これを基に新たなアノテーションが機械的もしくはアナログ的に付与されます。機械的な処理の例として、形態素ごとの文法的なつながりを解析しデータ化する構文解析が挙げられます。具体的にどのようなものであるのか、ここでは割愛しますが、「日本語書き言葉均衡コーパス」や「日本語話し言葉コーパス」では、構文解析器を使って文の係り受け情報を付与していますので、興味のある方はドキュメントを読んでみると良いでしょう。

紙面の関係上、ここで示したアノテーションの例は 1 文だけですが、現在のコーパスはそれこそ数百万文～数億文が収録されおり、それぞれにこのようなアノテーションが付与されています。それくらいの分量になると、ことばの研究として文章の全体の傾向を観察したり、レジスターごとの語彙や文法の傾向を比較する対照研究が可能となります。更に超大規模のコーパスであれば、機械翻訳や人工知能の学習用データとして使用できるようになります。現在、人手であれば膨大な時間がかかるはずの言語データの整理とアノテーションが、形態素解析をはじめとする技術によって、短期間かつ少ない労力で実行できるようになりました。この技術は、ことばの研究と機械によることばの習得の基盤として必要不可欠なものとなっているのです。

言語学からみた機械のことば

現在、大量の文が多くの技術によって処理され、機械翻訳や人工知能の技術に利用されてきたおかげで、人工知能のタスクや翻訳の精度は大幅に向上してきましたが、決してそこに課題がない訳ではありません。機械翻訳の場合、「翻訳」という行為の本質を振り返ったときに機械翻訳には原理的にできないことがありますし、元言語から目的言語への翻訳をするにあたって言語データだけでは補いきれない言語学的な課題があります。人工知能の場合、人間が発したことばを適切に受け取り、理解することや、人間と適切なコミュニケーションを取ること、人工知能自体の経験を語ることについて課題が残っています。ここでは、機械がことばを理解し話すにあたって残されている課題について、いくつか紹介したいと思います。

言語学からみた機械翻訳の課題

先ほども見たように、現代の機械翻訳は確かに精度がきわめて高く、一見したところ、人間のことばを十分に理解しているように見えますが、現代の機械翻訳で適切に翻訳できない文には実際どのようなものがあるでしょうか。景浦 (2017: 931) の論文で挙げられた英語の文を例に見ていきましょう。

“Human” consists of 5 letters. →「人間」は 5 文字で構成されています。

上の英文を GoogleNMT で日本語に訳すると、下の文のように出力されます（2020 年 6 月時点）。本来の文意を踏まえれば “Human” はそのまま “Human” として残されなければなりませんが、「人間」と翻訳されているために、文に矛盾が生じてしまっています。上の翻訳はたしかに元言語と目的言語のそれぞれの単語の対応関係を（言語学的な側面で言えば）適切に表しているでしょうが、元の文章で表される「言説」を目的言語でも適切に表せるかというと、必ずしもそうではないのです。

同じ論文の中から、もう一つ例を挙げましょう。下の文は人種差別撤廃条約の英文とそれをGoogleNMT で日本語に訳したものです。

> States Parties shall assure to everyone within their jurisdiction effective protection and remedies, through the competent national tribunals and other State institutions, against any acts of racial discrimination which violate his human rights and fundamental freedoms contrary to this Convention, as well as the right to seek from such tribunals just and adequate reparation or satisfaction for any damage suffered as a result of such discrimination.
> →締約国は、管轄区域内のすべての人に、管轄の国内裁判所および他の州の機関を通じて、この条約に反する人権および基本的自由に違反する人種差別行為ならびに求められる権利に対して、効果的な保護および救済策を保証するものとします。 そのような法廷から、そのような差別の結果として被った損害に対する正当かつ適切な賠償または満足から。

機械翻訳によって出力された日本語文は後半が途切れたような内容になってしまっていますが、問題はそれだけではありません。日本政府はこの人種差別撤廃条約を部分的に批准しているので、この条約に公式な日本語訳が存在するはずですが、機械翻訳はプログラムが学習したこれまでのデータに基づいて翻訳をおこなうだけで、その公式な日本語訳を引用することはありません。

機械翻訳に残されているその他の課題として、対訳コーパスに出現しない固有名詞や新語・未知語の処理があります。たとえばゲームのタイトルは固有名詞の一例です。2020 年 3 月に任天堂が「あつまれ　どうぶつの森」というコミュニケーションゲームを発売し、2020 年 6 月現在、世界中で大きな反響を起こしていますが、このゲームのタイトルを GoogleNMT に入力しても英語版のタイトル “Animal Crossing: New Horizons” は出力されません[2]。これはゲームの名前が固有名詞であるだけでなく、ゲームに関する記事や比較的新しいトピックに関する記事の翻訳データが対訳コーパス内に十分存在していないことが理由として挙げられます。新語・未知語の例としては若者言葉があります。試しに 2020 年現在の若者言葉「詰んだ（策が尽きたこと）」「草（たいへん可笑しく笑える様子）」を確認しましたが、これらの語は適切に翻訳されていません。これらの新語は、機械翻訳が単体で対処するにはなかなか難しい課題と言えるでしょう。現在、GoogleNMT はこの課題への対処として、利用者が不自然な翻訳を指摘し、適切な語を入力できる機能を搭載しました。この指摘によって GoogleNMT は対訳データを学習し、適切な翻訳に修正することができます。例えば、2010 年代の若者言葉「激おこぷんぷん丸」はこの機能によって、一定の適切さをもって翻訳されています。機械翻訳の課題を人が適切に補っている例と言えるでしょう。

それでは機械翻訳に対して、人間がある文章を別の言語へと翻訳する作業とはどのようなものでしょう？翻訳者がその文章の内容だけでなくその背景やニュアンス等を十分に理解している場合は、それらを十分に再現できるように目的言語の語や文法を構成するように試みるでしょう。またそこで新語・未知語、固有名詞を見つけたとしても、情報収集と推論によって可能なかぎり適切な訳語を探し出すでしょう。しかし、翻訳者がその文章について十分に理解していない場合

[2] 試しに他のゲームタイトルも入力してみましたが、適切に翻訳されているゲームタイトルはかなり少なそうです。人気になっているゲームがそれぞれの国によって異なることも、翻訳記事の量に影響しているかもしれません。

は、文章を構成する語を、辞書を片手に一つひとつ訳したり、文章の文法構造をそのまま目的言語の構造に当てはめていくように試みるでしょう。現在、高い精度を保っている機械翻訳は、後者のような翻訳を実現していると言えますが、前者のような（人間にとっても難易度の高い）翻訳まで実現するには、解決すべき課題が残されていると言えます。

世界の経験とことばの獲得

では、人工知能やロボットがことばをあつかうときの課題には何があるでしょう？それを直接お話する前に、Deb Roy の TED 講演「はじめて言えた時（The Birth of a Word)」を紹介し、人間と人工知能やロボットでことばの獲得の仕方がどのように違うのかを見ていきたいと思います。この講演では、Roy 氏が自分の子供と（Roy 氏自身を含む）居住者の自宅内のやりとりの映像と音声を、子供が生まれてから 3 年間収録し続けたデータベースを構築し、そこから自宅内の人々のコミュニケーションの解析を通して、子供がどこでどのような語彙を学んでいくのかを紹介しています。収録は 1 日約 8〜10 時間にわたり、3 年間の収録データは 9 万時間分の映像と 14 万時間分の音声が含まれています。Roy 氏によるデータの中に無音状態のデータが含まれているのかどうかは定かではありませんが、日本国内で構築されている音声・映像付きのコーパスでも大きいもので 200 時間分なので、たいへん規模の大きいプロジェクトだったことが分かります。講演では、子供が水を指すために言う “ga-ga” という発音が次第に “water” というはっきりした発音へと変化していく様子を、まるでつぼみが花開く様子を早送り再生する映像のように聞かせてくれます。Roy 氏は子供と居住者の行動を細かく辿り、どこで、誰が、誰と、どのような場面で、何を発話したのかを観察・分析しました。その結果、たとえば “water” であればキッチンやバスルームの周辺で多く発話され、“bye”であれば玄関で多く発話される等、語と場に強い結びつきが存在することを具体的に明らかにしました。彼はこの語と場の結びつきの強さを地形図のような図（word-scape）として視覚化しています。

またこの講演ではもう一つ重要な指摘がなされています。それは、子供とコミュニケーションを取る大人たちも、子供の言語の習得段階に応じて文の複雑さを適切に調整しているという点です。Roy 氏は、大人が特定の単語を発した時の会話を収集し、その単語が使われている発話の相対的な長さがどのように変化するのかを調査しました。その結果、子供が単語を習得するまで大人の発話は最小限に短くなっていき、その子供が単語を習得した時点から大人は発話を少しずつ長くしていくという傾向が確認されました。これは子供の親と養育者に共通する傾向であり、大人は無意識のレベルで子供の単語の習得状況の把握と、その状況に合わせた適切な習得環境の構築をおこなっている、と Roy 氏は考察しています。

子供は日常生活の様々な経験を通して語を習得します。自分がいる場に誰が存在し、どのような出来事が起きたか、そこで自分が五感で何を感じたか、その場で誰がどのようなことばを発したか、というような経験一つひとつが少しずつ蓄積していき、それがある時子供の発話の中で特定の意味を持って立ち顕れます。また子供が順調にことばを習得できるように、大人たちも発話の複雑さを調整することをはじめとして、習得しやすい環境を整えてあげるという相互行為をおこなっています。人間が習得する語とその意味は記号的なものではなく、むしろ身体と環境の関わり合いや相互行為をとおして得られるものであり、一種の生々しさを持つものです。人間のことばの獲得は経験基盤的なプロセスを経ていると言えます。

一方、人間の身体を持たない人工知能やロボットの場合、ことばを外部の事物に結びつけたり、人間（もしくは他の機械）とのコミュニケーションを実現するかという点が大きな課題となります。人工知能やロボットがことばと外部の世界をどのように結びつけるかに関する課題はHarnad (1990) によってシンボルグラウンディング問題と呼ばれ、多くの研究者が解決を試みている問題です。人工知能やロボットが人間の子供と同じように経験基盤的な意味を獲得するには、まずその経験を受け取る感覚器官としてのセンサーを持つか、もしくはそのような情報を外部からデータとして得ることが必要となります。前者の場合、どのようにロボットの構造やセンサーをデザインし、どのような場でどのような経験をどれだけ長くさせるかが課題となります。一方後者の場合、大規模なコーパス等のデータベースを機械に学習させることで、外部の世界の情報を高い精度で認識させることが可能です。現在は画像検索を利用することで特定の語とその語に対応する画像を大量に学習し、具体的な事物や生物、人物等の姿とことばを結びつけることが可能です。しかし画像として表現することが難しい抽象的な単語や比喩的な表現に関する理解はきわめて難しいと指摘されています（川添　2017）。また機械が外部からデータを取得し学習することで、その機械自体の経験とデータで得られた情報が混同されてしまうという問題もあります。たとえば福岡で作られた機械に「あなたはどこから来たの？」と質問しても、その機械が学習したデータに似たような質問と回答（例: アメリカ生まれ）があると「私はアメリカで生まれました」と回答してしまうというような例が挙げられます。

またことばはただ情報を伝えるだけでなく、その情報がどれだけ確かなものかという見込みや情報に対する肯定的・否定的感情、情報を伝えることで相手にしてほしいこと等も伝える重要な機能を持っており、これらの機能は言語学の語用論の分野等で研究されています。人間はこの語用論的な機能を若い頃から理解し習得しますが、機械にとっては抽象的な概念と同様に習得が難しい部類の機能です。たとえば「花を送ってくれたなんて、彼は良い人じゃないの。」という文は基本的に彼が「良い人である」意味として解釈されますが、機械翻訳では彼は「良い人ではない」という真逆の意味で翻訳されてしまいます。また「（火事のときには）建物の外に出て火の元に近づかない。」という文（Cf. 川添　2017:223）を読んだとき、文中の「ない」は通常 2 つの行動（建物の外に出る、火の元に近づく）のうち後者のみを否定するものであると我々は解釈しますが、機械にとっては、火災等の非常時の対策に関する知識がない場合、前者と後者両方を否定する文として解釈すべきなのか否かを判断できず、曖昧性が残ってしまいます。人間にとって簡単に判断できることは、実は機械にとっては簡単ではない場合があるのです。

一つ興味深いプロジェクトを紹介しましょう。国立情報学研究所は 2011 年から「ロボットは東大に入れるか」プロジェクト（略: 東ロボ）を発足し、人工知能分野を再統合し、2016 年までに大学入試センター試験で高得点をマークすること、また 2021 年度に東京大学入試を突破すること を目標に研究活動を進めています（プロジェクトホームページより）。このプロジェクトは 2016 年の時点で一定の成果を収めていますが、注目したいのはこのプロジェクトの後に同じく国立情報学研究所で発足した「ロボットは井戸端会議に入れるか」プロジェクト（略: 井戸ロボ）です。井戸ロボでは井戸端会議に入れるロボットの構想を通して、人間の社会的なコミュニケーションがどのように行われているのかを探る研究が行われています。東ロボが東大に合格するロボットの開発をとおして人間の知能や知性にアプローチしている一方、井戸ロボは人間の社会性にアプローチしているという対比がおこなわれています（西田　2013）。機械にとって人間同士の会話は実に曖昧で正解を得るのが難しいものですが、人間がそれを日常的に流暢におこなって

います。機械が人間とそっくりなコミュニケーションを実現するには、まだ多くの課題が残されているのです。

まとめ

さて、この授業では現在の機械に組み込まれている、ことばに関する高度な技術の紹介を通して、現代の技術がどれだけ人間のことばに迫っているのかを見てきました。その後、その技術を支える言語データであるコーパスを紹介し、その構造や規模の大きさや多様な形態を通して、コーパスの科学的・学術的な意義についてお伝えしてきました。最後に、人間のことばの実態を探究する言語学の視点から機械のあつかうことばにどのような課題が残っているのかを、人間のことばの実態との比較を通して考えました。

ことばを話す人間の実態を見直すと、機械が人間そっくりにことばを話したとしても、機械そのものが人間そっくりに思考し、ことばを使っているとはかならずしも言えません。しかし今の機械に残されている課題を解決しようと、研究者は日々さまざまなアプローチを試みています。この課題発見と解決のサイクルを通して、ことばに関する技術は向上すると同時に、人間がことばを使うにあたって必要不可欠な能力が何であるのかを再確認することでしょう。ことばをあつかう機械にまた新しい進展が起きれば、ぜひそれはこの授業の最初に紹介したいところです。

各年度のレポート課題と解説

この授業では毎度 2 つのテーマを設定し、それに基づいて科目の最後にレポートを提出してもらいました。以下では、各年度に出題したレポートのテーマを紹介したいと思います。いずれのテーマでも日常的なことばに意識を向け、ことばに対して科学的・創造的なアプローチの提案を求めるものだと思います。興味がある方には、どのようなレポートを書けそうか、ぜひ考えてもらえたらと思います。

2017 年度

目的に応じたコーパスの構築とその言語データの理解

> あなたはおもちゃの開発者で、子供の情操と知性を育むような、ぬいぐるみ型のロボット※の開発を目指し、適切な言語データを探しているとします。開発には「誰による」「どのような内容の」言語データ（コーパス）が必要か、その理由も含めて論じてください。
>
> ※対象年齢は 3〜6 歳とする。

この授業の中で Deb Roy による子供の言語獲得の研究をとおして、子供を取り巻く環境とそこにあふれることばのやり取りの重要性を学びました。ここではあえてロボットに人間のコミュニケーションを再現させるという課題を設定することで、言語獲得に強く関わる環境やことばの特徴を理解することを目的としました。ロボットが果たして子供の言語獲得に寄与しうるのか、また寄与しうるとしたらどのような点で寄与するのか、という可能性を学生が模索することを期待しました。

身近なコミュニケーションの言語学的理解と応用可能性

> 今、あなたの手元に「あなた」と「あなたが最も長く時間を過ごした 5 人の人物」による会話データ（音声・非音声を問わない）があるとします。そのデータの構成（主な会話形式・内容、特徴、各話者の特徴、各話者が占める会話時間のおおよその割合、等）を説明したうえで、そのデータの解析・処理によってどのような領域・課題への応用が期待されるかを論じてください。

この課題は自分の日常的なことばのやり取りを、コーパスのように構造化されたデータとして捉え、ことばの様々な側面が言語学的に重要であるという認識を持つことを目的としました。また「最も長く時間を過ごした 5 人」との会話を通して自分自身が彼らから（言語・非言語を問わず）どのような影響を受けているのかを考え直すきっかけとなればと思い、この課題を設定しました。

2018 年度

コーパスの目的と構築方法の理解

> **(i)** 現存するコーパスを 1 つ取り上げ、その（a）構築の目的、（b）対象のレジスター、および（c）内部の構造を簡単に説明してください。
> **(ii)** あなたは「九州大学学生コーパス」を構築することになりました。そのコーパスについて、前項 (i) の (a)～(c) にあたるものを自らデザインしてください。

この課題では 2 つの目的を設定しました。1 つ目は、コーパスの具体的な構造に関する資料を読むことで、コーパスの文章が目的に応じて語や形態素等の細かい単位に分割されている点や、その単位にアノテーションが加えられている点、特定のレジスターに絞って言語資料が収集されている点に関する理解を深めることです。2 つ目は、同じような視点から身近なことばのやり取りを見て、自分のことばに学術的な意義を見いだすことです。この 2 つの目的から、ことばを有用なデータへと変える手続きについて理解を深めることを期待しました。

非字義的な意味の理解

> あなたが最近、日常でおこなった会話を 2 つ挙げ、その 2 つの会話に共通する「含意」あるいは「非字義的な意味」を説明してください。また、その意味について、認知言語学もしくは語用論の用語を 1 つ以上挙げて考察してください。

この課題では、機械にとって難しい推論を必要とする会話がどれだけ日常的にあふれているかを理解し、言語学的な用語について一定の理解を深められることを目的としました。また、自分が日常的におこなっている会話を注意深く観察することを通してことばへの興味や関心を持ってもらえたらと期待して出題しました。

参考文献

麻生英樹・安田宗樹・前田真一・岡野原大輔・岡谷貴之・久保陽太郎・ポレガラ ダヌシカ (2015)『深層学習』 人工知能学会（監修），神嶌敏弘（編） 東京: 近代科学社.
原 倫太郎・原 游 (2008)『背面ストライプの浦島太郎 ～日本昔話 Remix2～ ～日本昔話 Remix～』 東京: マガジンハウス.
Harnad, Stevan. (1990) “The Symbol Grounding Problem,” Physica D Vol. 42: 335-346.
景浦 峡 (2017)「改めて、翻訳とは何か？: Google NMT が使える時代に」『言語処理学会 第 23 回年次大会 発表論文集』 pp. 931-934. 言語処理学会.
川添 愛 (2017)『働きたくないイタチと言葉がわかるロボット: 人工知能から考える「人と言葉」』 東京: 朝日出版社.
西田豊明 (2013)「特集 ロボットは井戸端会議に入れるか」 『NII Today』 第 62 巻, pp. 2-3, 国立情報学研究所.
奥村 学 (2010)『自然言語処理の基礎』 東京: コロナ社.
スルダノヴィッチ イレーナ・スホメル ヴィット・小木曽智信・キルガリフ アダム (2013)「百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング」『第 3 回コーパス日本語学ワークショップ 発表予稿集』 pp. 229-238. 国立国語研究所.
高村大也 (2010)『言語処理のための機械学習入門』奥村 学（監修） 東京: コロナ社.
友利 涼・亀甲博貴・二宮 崇・森 信介・鶴岡 慶雅 (2017)「シンボルグラウンディングによる分野特有の単語分割の精度向上」 『自然言語処理』 第 24 巻 3 号, pp. 447-461, 言語処理学会.
塚田 元・渡辺太郎・鈴木 潤・永田昌明・磯崎秀樹 (2007)「統計的機械翻訳」『NTT 技術ジャーナル』 第 6 巻, pp. 23-25. 日本電信電話株式会社.
土屋智行 (2019)「言語の慣習性を中心とした言語研究の手法と展開」 日本言語教育 ICT 学会 2019 年研究大会 基調講演, 2019 年 9 月 7 日, 久留米大学.
Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. and Klingner, J., (2016) “Google’s neural machine translation system: Bridging the gap between human and machine translation.” arXiv preprint arXiv:1609.08144.
山村ひろみ（編） (2018)『現代ロマンス諸語におけるテンス・アスペクト体系の対照研究』科研費報告書 CDR, 九州大学.

資料

Google Developers (2018) Keynote (Google I/O ’18): https://www.youtube.com/watch?v=ogfYd705cRs
Google Translate: https://translate.google.com/
国立国語研究所 (2006)『日本語話し言葉コーパスの構築法』
国立国語研究所 (2015) 『「現代日本語書き言葉均衡コーパス」利用の手引 第 1.1 版』 https://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html
国立国語研究所 (2017) UniDic ホームページ: https://unidic.ninjal.ac.jp/
MeCab: Yet Another Part-of-Speech and Morphological Analyzer: https://taku910.github.io/mecab/

ロボットは東大に入れるかプロジェクトホームページ: https://21robot.org/index.html
TED (2011) “Deb Roy: The birth of a word”:
https://www.youtube.com/watch?v=RE4ce4mexrU