挑戦する研究者たち
CHALLENGERS

永田昌明

NTTコミュニケーション科学基礎研究所
上席特別研究員

DIFFERENTはほめ言葉である．未来を論理的に語る訓練を心掛けよう

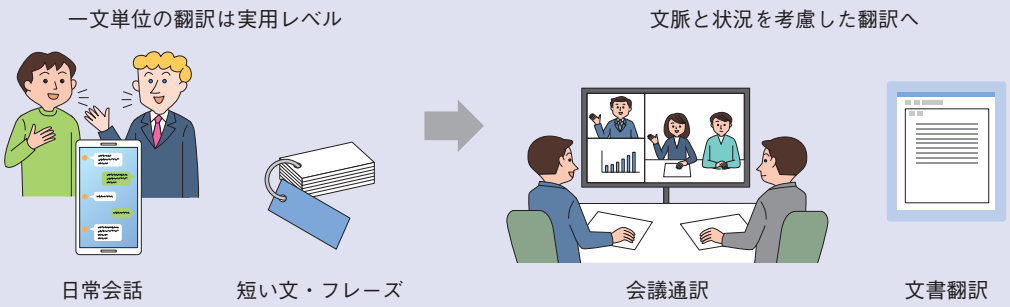
近年，ニューラルネットワークを用いた機械翻訳の研究は急激に進歩し，機械翻訳の精度は大幅に向上しました．Webサイトやスマートフォンのアプリによる翻訳が普及してきている中，その精度にはまだまだ課題が残されています．この精度向上をめざして研究者は文脈や状況，文化等をも反映したさらに高度な対訳技術研究に勤しんでいます．今回は，文脈・状況に基づくニューラル機械翻訳を追究する永田昌明NTTコミュニケーション科学基礎研究所　上席特別研究員に，研究の進捗と研究者としてのあり方について伺いました．

自然言語処理のトップレベルの国際会議で高評価

現在，手掛けている研究内容について教えてください．

前回お話させていただいた2013年から一貫して，ある言語を別の言語に翻訳する技術を追究していますが，対象は単語の対訳と文法をベースとした統計的機械翻訳から文脈・状況に基づくニューラル機械翻訳へと変化しました（**図1**）．2014年ごろからAI（人工知能）技術が急激に進歩し，囲碁の勝負でAIが人間に勝ち，難しいとされていた音声認識が可能となるばかりかその質が劇的に向上する等，AI関連の各研究分野に大きな変化が現れてきました．こうした中，私が手掛けていた機械翻訳の分野においても，2016年ごろに翻訳の精度が急速に向上し，ヒューマンパリティ，つまり人間並みの翻訳ができるようになりました．これまで不可能だと思われていたことが突然できる時代がやってきたのです．このような時代の流れを受けて，今後の研究テーマを模索する中で，前後の文により翻訳文の文意が変わってくるといった課題が山積していることに気付き，この解決をめざして文脈・状況に基づくニューラル機械翻訳というテーマにたどり着きました．

Webサイトやスマートフォンのアプリによる翻訳機能を使うと，その精度は高く高校生の宿題程度の英作文なら機械翻訳のほうが上手ではないかと思えるほどです．私は



ニューラルネット（深層学習）により翻訳精度は大幅に向上し，一文単位の翻訳は実用レベルになった．今後は文脈と状況の考慮が課題．

図１　機械翻訳の現在の到達点と次の目標

貸自転車屋さんで店員と外国人がスマートフォンの翻訳機能を使って交渉している場面に遭遇し，そのやり取りを見ながら「すごいな，翻訳技術もここまで来たか！」と感心しました．一方で，２人の間の会話は円滑とはいえるものではありませんでした．それは，微妙な言い回しや，ニュアンスは翻訳できていないからでした．

これは，ニューラル機械翻訳の少し困った特徴で，文章そのものとしては母語話者並みに流ちょうな訳文を生成する反面，訳文が原文の意味を忠実に再現しないことがあるためです．従来の機械翻訳システムは文を基本的な入力の単位としていたので，たとえ１つずつの文の翻訳精度が人間に匹敵したとしても，文書や会話のような複数の文から構成されるテキストを翻訳すると，文脈や状況を考慮していないために照応関係が一致しない，訳語に一貫性がないという問題が生じてしまうのです．

このような現在の技術をさらに進化させるためにご研究に臨まれているのですね．

これに対応するための課題は３つあります．１番目は日本語の会話では往々にして主語や目的語が省略されますが，その省略された主語や目的語を文脈から判断して翻訳すること．２番目は会話に登場してくる人物が誰かを反映した翻訳にすること．これには例えば「脳外科医」には男女両方いるにもかかわらず，つい男性を連想してしまうようなジェンダーバイアスの問題も含まれます．そして３番目は１つの単語が保有する複数の意味や表現から適切な意味や表現を見つけ出して反映することです．

文脈や状況を考慮した翻訳の例として日本語を英語に翻訳する場合を考えます．日本語の第１文に依存して，日本語の第２文の英語訳が決定される仕組みです（**図２**）．

「申し訳ありませんが，先生は午後，少し遅れているのです．診察するまでに20分ほどかかると思います」という日本語を英訳すると，“I am afraid that the doctor is running a bit late this afternoon. It might be about 20mins before he can see you.”が正答として提示されます．しかし，日本語の第２文において「先生」という主語が省略されているため，“It might be about 20mins before we can see you.”のように，ドクター（三人称単数）に該当するはずの主語が「私たち」になってしまうことがあるのです．さらに，これには，ジェンダーバイアスが存在します．英語では代名詞を選択する際には性別を特定しなければいけません．しかし，機械は「ドクターは男性である」という認識の下，“he”と翻訳してしまうのです．

このように日本語から英語に翻訳する際，日本語において省略されてしまう主語や目的語をうまく解釈しないと，

	原言語	目的言語
先行文	申し訳ありませんが，先生は午後少し遅れているんです．	I’ m afraid that the doctor is running a bit late this afternoon.
入力文	診察するまでに20分ほどかかると思います．	正解: It might be about 20 minutes before he can see you. 誤り: It might be about 20 minutes before we can see you.

	原言語	目的言語
先行文	昨日，渋谷へ行った．	I went to Shibuya yesterday.
入力文	すごい人だった	正解: There were a lot of people. 誤り: He was a great man.

さらに英語の代名詞に性別の区別があるため「先生」や「人」の訳語にジェンダーバイアスの問題が生じる．

図 2　文脈の考慮が必要な翻訳の例

意味が変化してしまうことに加えて，訳語が文脈によって変化してしまうのです．例えば「昨日，渋谷に行った．すごい人だった」という日本語における「すごい」のニュアンスは「渋谷に行った」を受ければ「人出の多さ」だろうと考え，“There was a lot of people.”となるでしょう．しかし，文脈を反映せずに別の意味で解釈されると，“He was a great man.”等と，文としては間違いありませんが前の文に呼応しない翻訳文になってしまうのです．

さらに，どんなに翻訳精度を高めても完璧な翻訳はできませんから，誤訳にどうアラームを鳴らすかも重要です．例えば誤訳が発見されたとき，統計的機械翻訳の場合，翻訳された文章が元の文章のどこにあたるかを調べたいときには，カーソルを文章に当てるとその部分を示すことができました．ところがニューラル機械翻訳は文全体を解釈して意味ベクトルを作成し，そのベクトルに従って翻訳がなされるので対応する部分を具体的に特定できないのです．

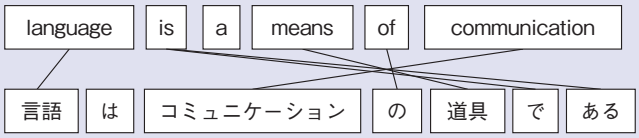
こうした状況をかんがみて，単語対応，文対応といわれる，入力した文章と出力した文章の対象を具体的に示すことを目標に研究を進めました．そして，2020年に自然言語処理・計算言語学の分野での世界最大の国際学会の１つであるEMNLP(Empirical Methods in Natural Language Processing)において，文脈を反映してお互いに翻訳になっている部分とそうでない部分を判別できることを発表しました（**図 3**）．この取り組みは先駆的で精度が著しく高く，画期的だと評価されました．

人と違うことをして，できるだけ物議を醸す

素晴らしい研究成果を上げられたのですね．研究活動をするうえでは何を大切にしていらっしゃいますか．

私は人と違うことをすること，できるだけ（いい意味で）物議を醸し出すことをしたいと心掛けています．ニューラル機械翻訳において単語対応ができないのはとても不便であるなどの問題意識を持っており，この問題を解消するため，他の人とは違う方法を試して模索した結果，偶然ではありますが，「言語が違っていても，ニューラルネットに，ある単語と別の単語が意味的に近い存在であるかを学習させるのは，実は簡単で，300文程度の単語対応の正解データがあれば高い精度で単語対応を求められる」ことを発見したのです．

発見に至るプロセスは，まずmultilingual BERT (Bidirectional Encoder Representations from Transformers)*で学習した単語の意味ベクトルを眺めて



language is a means of communication
Q. 言語
A. Language
Q. は
A. No_Answer
Q. コミュニケーション
A. communication
…

日本語文の単語（スパン）に対して英語文で対訳となる単語列（スパン）を求める．
双方のスパンの意味ベクトル表現から対応関係を決める．

図 3　言語横断スパン予測に基づく単語対応の例

みたところ，多言語間の対訳文をデータベース化した，いわゆる対訳コーパスを学習に全く使用していないにもかかわらず，でき上がったモデルの中では，実は似たような単語はベクトル空間上の割と近い場所に存在している状態にあることに気付いたことです[1]．

その理由は言語が違っても数字やアルファベット等で表記された部分は共通であることや，日本語と中国語も漢字を使用することで重なるように，近い言語どうしであれば何らかの共通部分があるからです．それを眺めたときに，共通する表記や言葉がピボットとなってそれを中心に各言語が独自の世界を広げるという状況が起きているのかもしれないと考えました．この着想からニューラルネットにある単語と別の単語が意味的に近い存在であるかを学習させるのは実はとても簡単で，300文程度の単語対応の正解データがあれば高い精度で単語対応を求められることを導き出したのです．

試行錯誤を繰り返して「こんなやり方もあるかもしれないな」と何気なく取り組んだことが画期的な発見に結びついて驚きましたが，これも人と違うことをして，物議を醸すことをしようという心掛けによるものです．

こうした考えに至ったのは30代前半に米国で勉強させていただいた経験に由来するのかもしれません．私は２度渡米しています．１度目はCMU（Carnegie Mellon University）での経験です．彼らは誉め言葉として“He is a different.”と表現していたのを聞き，“different”は誉め言葉だと知りました．また，２度目のAT&Tでのスーパーバイザであるケネス・チャーチ博士も，「私は人と違うことをして，物議を醸すものをねらう」と話しており，私は研究者としてその姿勢に深く共感しました．この経験をきっかけに「違う」ことは良いことなのだと価値観が変わりました．

また，２度の渡米を経てサーベイの重要性を学びました．人と違うことをするには人と同じ部分を理解していなくてはなりませんから，自らの研究がどのポジションにあるか，

＊ BERT：Googleによって開発された，Transformerと呼ばれる深層学習法を用いて大量のテキストから事前に学習された言語の意味表現モデル．テキスト分類など目的別の正解データを用いてさらに学習すると高い精度が得られる．Multilingual BERTは104言語のテキストから1つのモデルを学習したもの．

何が分かっていないか等のサーベイには今も真摯に取り組んでいます．

米国での経験は研究者としての歩みに大きく影響したのですね．

CMUでは研究者の心構えも教えていただきました．米国の大学の新学期にあたる夏から秋にかけて，研究についてのガイダンスが開かれていました．そのときに「研究とは人間の知能の最前線にあって未解明の物事に英知を持って臨む」と学びました．研究とは何か，何に貢献すべきか，といった基礎となる概念を明確に教えていただく機会はそれまでの研究者生活においては少なかったので，早い段階でそれ知ることができてラッキーだったと感じています．また，米国では博士課程において，当時から研究の方法や成果，到達度等，博士号への評価指標が具体的に示されていましたから，研究への取り組み方や目標設定をしやすかったのを覚えています．

NTTにも「テーマ企画」と呼ばれる非常に良い伝統があります．これは入社1，2年目の研究者の研究企画発表会です．登壇者は3，4年間の研究計画を理論立てて発表します．これは米国の大学院の“Thesis proposal（論文計画書）”ととても良く似ています．どちらにおいても，自らが研究したこと，いわゆる過去について語ることは，研究者として経験を重ねることでおのずと身についてくることですが，計画を語る，いわゆる未来を論理的に語ろうとすることは訓練が必要です．その機会を若い研究者にシステムとして与えるのは非常に良いことだと思います．こうした活動から自らの立脚点を理解して，使命を明確にすることが人と違う研究につながり，物議を醸す研究成果を上げることにつながると考えます．

こうした考えに基づいて，私は常に自分の研究を客観的に眺め，新しいことに取り組むようにしています．最近は言語の外側にある現実世界の「状況」をニューラル機械翻訳に反映するためにグラフニューラルネットワークの勉強を始めました．グラフニューラルネットワークはさまざまなところで実用化されています．例えば，最近実用化されたのはGoogleマップで到着時間を予測する機能です．道路のすべての交差点と道路のリンクの間，人の移動の計測をして，A地点からB地点へ移動するのにかかる時間を時刻ごとに予想します．また，Uber Eatsのレコメンデーション機能も同様に，注文者と注文した物等を組み合わせて当事者が次回に注文する物を予測しています．このグラフニューラルネットワークのように言語化されない現実世界の「状況」に関する情報を反映することで，翻訳の精度をさらに向上させることができると考えています．

前代未聞の「辞めます」宣言

これまでの研究活動を振り返って印象に残っていること，そして今後はどのようなことに取り組むのかを教えてください．

2016年の夏に口にした「統計的機械翻訳辞めます」宣言が印象に残っています．30年ほど機械翻訳について研究してきましたが，ニューラル機械翻訳の登場による技術変化によって機械翻訳の研究は大きく変化し，過去の研究が帳消しになるほどのインパクトでした．これにより，ニューラル機械翻訳の分野の研究者は一斉にスタートラインに立つことになりました．こうした状況の中，私は研究方針を報告する会議で「統計的機械翻訳はニューラル機械翻訳には勝てません，だから辞めます」と宣言したのです．今から思うと前代未聞だっただろうなと思います．

それまでチームで取り組んでいた統計的機械翻訳の研究において，年長でリーダーだった私は，第一線から退いたような感覚でした．一方で，技術革新に伴ってそれまでのチームから個人で研究できるようになり，1人の研究者として復活することができました．そして，管理職としてチームを率いる立場の年齢になって，再度第一著者(first author)として論文発表することができたことはとても感慨深いものがあります．

今後は研究の土台になるような何かを残したいですね．どんな研究をするうえでも，データは非常に重要な存在です．機械翻訳を研究していくうえでは，土台が対訳コーパ

スや翻訳データベースではないかと思います．多言語間の対訳文をデータベース化したものを対訳コーパスと呼びますが，これを作成するのに必要なのがデータ源なのです．初期には国際機関の文書をデータ源としていたのですが，現在の機械翻訳の典型例といわれる，Google翻訳はWebをデータ源としています．一方で，Googleのような１つの企業が独占的に情報を蓄積するのは良い状況とはいえないという意見もあり，Webをクロールしてデータを一般に提供することを目的とするNPOであるCommon CrawlのWebアーカイブから，大規模に対訳文を抽出して対訳コーパスを作成する試みがいくつか行われています．

このCommon Crawlを利用したヨーロッパ言語―英語間の大規模な対訳コーパスを作成するプロジェクトParaCrawlでは，4000万文を集積した段階で機械翻訳の精度が人間と同等になったといいます．日本語はそこまで蓄積がありませんから，このままだと機械翻訳業界は欧州においていかれてしまいます．このため私はNTTが作成した日英対訳コーパスJParaCrawlを充実させ，日本語の翻訳研究の発展に尽力したいと考えています．

過去に定年退官間近のある大学教授が学生たちを動員して，後にtatoebaと呼ばれる大規模な翻訳データベース製作に臨まれていました[2]．実は当時，「定年間近でこんなことを始める意味は何だろう」と，その先生の熱い思いを理解できなかったのです．しかし，自分自身がその年齢に近づいた今はその気持ちが分かります．多くの人がアクセスできるデータベースを作成する活動には引き続きかかわっていきたいと考えています．

若い研究者へアドバイスをお願いできますでしょうか．

研究者にとっては，自分の興味を持ったテーマを長期にわたって追求するほど有利な状況になります．そのためには，他者に分かりやすく説明できる力を身につけること，そして，定期的に自分の研究と向き合うことの２つを大切にしてほしいと考えています．

定期的に自分の研究と向き合うことは，研究企画書を書くことでもそのスキルを養えます．NTTでは毎年，１年間の研究計画を提出します．その際に１年前の研究計画に目を通すことになるのですが，ベストエフォートの計画とはいえ，いかに自分の予測が当たらないかを目の当たりにすることになるのです．ところが，何度か繰り返すうちに実行できたことや予測どおりになったこと，あるいは反対に実行できないこと，予測に反したことが分かるようになります．経験を重ねることでより現実的な予測に基づいた研究計画を立てられるようになり，第三者にそれを分かりやすく伝えるためのスキルも養えます．

さらに，長期的に取り組むためには興味のあるテーマでなければなりません．私自身，自分の将来の見通しが立っていたかというとそうでもありません．かつては５年，10年を見据えて動いてはいましたが，リアリティがあるのは最初の１，２年かもしれません．しかも，変化の激しい今の時代において５年，10年先の未来を予測することは困難極まります．こうした状況においても，想像できる範囲で，想像した未来において自分はどうありたいかを描いてみることが大切だと思います．

若い研究者の皆さん，研究者ほど現役が長い仕事は他にないかもしれません．自分が長く付き合えるテーマを見つけて研究活動を続けていってください．

■参考文献

⑴ https://www.aclweb.org/anthology/2020.emnlp-main.41.pdf
⑵ https://tatoeba.org/jpn/