特許文献とAI翻訳の勃興

Patent documents and the rise of AI translation

国立研究開発法人情報通信研究機構（NICT）フェロー／アジア太平洋機械翻訳協会（AAMT）会長／
国際機械翻訳協会（IAMT）副会長

隅田　英一郎

1982 年電気通信大学大学院電子計算機学専攻修了。1999 年京都大学大学院博士（工学）。2014 年～ 2015 年言語処理学会会長、2015 年～現在、日本翻訳連盟理事、2018 年～現在、アジア太平洋機械翻訳協会会長。日本アイ・ビー・エム、国際電気通信基礎技術研究所を経て、情報通信研究機構で、2020 年をゴールとする音声翻訳の国家プロジェクト「グローバルコミュニケーション計画」を推進。情報処理学会喜安記念業績賞、文部科学省科学技術分野の文部科学大臣表彰科学技術賞、内閣府産学官連携功労者表彰総務大臣賞ほか受賞。

eiichiro.sumita@nict.go.jp

1 はじめに

AI 翻訳[1] は翻訳精度が高く評価されて、短期間に特許文献をはじめとして様々な分野で実用化された。

本稿では、「言葉の壁」の解消の緊急性と機械翻訳を利用した解消策について述べ、日本語と多言語の翻訳で高精度を実現した AI 翻訳の基本と旧技術にない AI 翻訳の特徴を概説し、特許庁や総務省等の官の様々な施策に加え、産官の連携である翻訳バンクについて述べた後、最後に AI 翻訳に戻り、今後の見通しで結ぶ。

2 「言葉の壁」とその解消策

急速に進むグローバル化によって、特に、次の三つの点に起因した「言葉の壁」の解消策が日本の喫緊の課題になっている。

中国特許の激増

広く知られているように、中国は、特許出願数を急増させており、その勢いは衰えを全く見せない。GDP 第 2 位の中国への輸出は、日本の各企業にとって魅力的であり、輸出に際しては中国国内の出願に抵触しないことが必須となる。

しかしながら、膨大な中国語特許を翻訳するだけの人数の中日翻訳者は集めることは現実的には不可能である。

観光客の激増

観光は新しい産業として期待されている。2014 年の 10 月に免税品枠が、それまで比較的高価な家電、着物、バッグなど限定されていたところから、食品、飲料、薬品、化粧品などまで拡張された。これが奏功し、ドラッグストアやコンビニなどに沢山の外国人が日本製品を購入するため立ち寄るようになった。また、クルーズ船の活発な誘致活動が実り、全国の地方都市に、年に数十回、数千人乗る大きな船が海外から寄港するようになった。多数の外国人が船から降りて、地元のお店に行って、買い物をしたり、食事をしたりする。その効果もあって、2015 年には、年間約 2000 万人の外国人が来日した。これを受けて政府は 2020 年の目標を 4000 万人に上方修正し、2030 年の目標を 6000 万人にした。既に 2018 年に 3000 万人を越えたところである。外国人観光客が一番多い国フランスには、年間 8000 万人以上が訪れることを考えると、観光資源が豊富にある日本の観光産業の伸び代はまだまだ大きい。

訪日観光客の約 8 割がアジアからの観光客。アジアの人たちが日本語を話せるとは考えられないし、英語を話せるアジアの人は少ないので、アジア言語への対策は重要であるが、アジア諸語の通訳者を必要な人数集める

1　ニューラル翻訳（NMT）とも呼ばれる。

ことは、ほぼ不可能である。

外国人就労者の激増

少子化に起因して日本は人手不足である。飲食店やコンビニエンスストアでは、外国人店員を見ない日はなく、日本社会を支える働き手としての存在感が年々高まっている。また、2019年4月より、新・在留資格である「特定技能」が新設された。建設業界など14業種が対象業種となった。ベトナム、インドネシア、ミャンマー、ネパール、カンボジア、モンゴル等幅広くアジア全体からやって来る。

人間の通訳者・翻訳者では全く足りない。

「言葉の壁」の解消策

英語以外の外国語の翻訳者・通訳者の数は非常に少なく、これを短期間に必要な人数まで引き上げることは不可能である。また、英語にしても、翻訳・通訳すべき情報の一部しか対応出来ていないと言われている。

ここで機械翻訳が登場する。

多言語化が容易であり、安く、速く、24時間365日利用可能である等メリットの多い機械翻訳は、AI翻訳の出現以前は、日英の翻訳精度が十分でなかったので、出番がなかった。ところが、AI翻訳が瞠目すべき精度改善を示してから、機械翻訳がメイン・プレーヤになった。

3 AI翻訳の早わかり

「AI翻訳の基本」

AI翻訳とは、①対訳データ[2]と②深層学習に基づく翻訳技術である。

対訳データを用いるというアイデアは長尾が1981年に提唱[3]し、その後、用例翻訳と統計翻訳という二つの少し異なる手法として大いに発展したが、日本語と英語のように文法が著しく違う言語対では、十分な精度が出せずに、近年、精度向上はプラトーに到達していた。

深層学習は昨今のAIの基本技術で、多層のニューラルネットで入出力データから変換を学習するものである。2012年のヒントン[4]が画像認識での成功を口火に、音声認識、碁などのゲームで華々しい成果を上げた。これを翻訳に適用し、最初の良い結果が出たのは2014年[5]で、2015年[6]に大きく改良され、その後もどんどんアルゴリズムは良くなっている。AI翻訳はその前提となる対訳データや自動評価などの研究リソースを再利用していることもあり、進歩が速くまだまだ天井が見える状況にはなっていない。

AI翻訳の欠点は桁違いの計算量にある。現在のAIでは、問題を積和計算に落とし込んで解くが、通常のCPUでは積和計算で十分な速度をだせない。このことから、CPUでなく処理能力は優れているが高額なGPGPUに依存しているところがAI翻訳の普及阻害要因である。

AI翻訳の利用方法

AI翻訳は既存の機械翻訳の周辺技術と融合されているので、様々な状況に合わせた利用が可能である。

例えば、自動翻訳サイト「みんなの自動翻訳」[7]サイトでも多数の方法が用意されている。テキストを入力して翻訳ボタンを押す。OFFICEのプラグインで翻訳する。TRADOS、MEMSOURCE、memoQなどの翻訳支援ツールから呼ぶ。WebAPIもある。

AI翻訳の分野専用化

基本となるアルゴリズムやハードウェアは研究者に任せておけばよいが、分野毎の対訳データによるAI翻訳の分野適応（アダプテーション）[8]、前・後処理・用語集

2 または、対訳コーパスとも呼ばれる。

3 1981年に発表されたが、論文「A Framework of Mechanical Translation between Japanese and English by Analogy Principle」」が公開されたのは少し後の1984年になる（http://www.mt-archive.info/Nagao-1984.pdf）。

4 https://wired.jp/2019/03/30/godfathers-ai-boom-win-computings-highest-honor/

5 Sutskever, Ilya; Vinyals, Oriol; Le, Quoc Viet (2014). "Sequence to sequence learning with neural networks". NIPS.

6 Thang Luong; Hieu Pham; Christopher D. Manning. (2015) Effective Approaches to Attention-based Neural Machine Translation. EMNLP.

7 https://mt-auto-minhon-mlt.ucri.jgn-x.jp/

8 M. Amin Farajian; Marco Turchi; Matteo Negri; Nicola Bertoldi; Marcello Federico. (2017) Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario. EACL .

の活用等は AI 翻訳の研究者以外が進めた方が効率的である。

4 特許庁と機械翻訳

特許庁は、極めて早い時期から、その事業の中で、機械翻訳を活用してきた。さらに、2007 年に、内山と特許庁は、文章の対から文の対を導出する内山のアルゴリズム[9]で、日本と米国への同時出願特許文献から世界で初めて特許の日英対訳コーパスを構築した。本対訳コーパスを利用した機械翻訳に関するワークショップを国際的会議 NTCIR の中で実施し、日英・英日の機械翻訳研究の進展に貢献した。その後、2014 年に特許庁と NICT は組織として共同研究を開始[10]し、数億文にのぼる対訳コーパスを共同で構築し、これに基づき統計翻訳から AI 翻訳まで機械翻訳の改良[11]に緊密に協力してきている。

この共同研究の成果である AI 翻訳・統計翻訳は、ルールベース翻訳と速度・精度の 2 つの観点から最適な形で統合したシステムとして実用化され、特許情報プラットフォーム（J-PlatPat）で公開され広く利用されている[12]。

5 総務省等の最近の動き

現在、日本政府は機械翻訳の研究開発に注力している。2014 年から、音声翻訳を 2020 年までに社会実装することを目指したグローバルコミュニケーション計画[13]が、総務省が音頭をとる形にてオールジャパンで実施されている。日本の各企業から研究者を NICT に糾合し、機械翻訳他に必要な機能の一つの優秀なソフトウエアとして作成し、これを各企業に技術移転し、各社の工夫を加えることによって、音声翻訳の市場を創出するものである。現在、計画の最終段階に到達しており、多種多様な製品が活発に上市され、100 億円を超える市場が出現した。

また、経済財政運営と改革の基本方針 2019、（いわゆる、骨太 2019）には、「深層学習による同時通訳を含む自動翻訳システムの開発・普及」と明記[14]されたところであり、政府の重点施策となっている。また、AI 戦略 2019（令和元年 6 月 11 日　統合イノベーション戦略推進会議決定）[15]でも自動翻訳の研究が重点化されている。また、外国人材の受入れ・共生のための総合的対応策（平成 30 年 12 月 25 日　外国人材の受入れ・共生に関する関係閣僚会議[16]において多言語翻訳アプリの活用が明記されている。

6 日本発の翻訳バンクの仕組

AI 翻訳の精度を上げるために対訳データの質と量が重要である。高品質なデータが大量にあればよいのであるが、そのような理想的な状態は簡単には実現できない。各開発組織の AI エンジンを比較すると分野、言語対等が違うときに、翻訳精度でみた順位が変動するのは各開発組織が有する対訳の相違に帰着すると想定しても大きな誤りにはならないだろう。

このような状況から脱却するために、世界で初めて、総務省と NICT は分野毎の対訳データを公的機関に集約する翻訳バンク[17]というスキームを 2017 年より開始

9　文書レベルの原文と翻訳から、文レベルの対訳対を抽出するアルゴリズムである。当アルゴリズムは、日本語文書と類似した英語文書を検索したあとで、その文書内での対訳文を抽出するものであり、必ずしも直訳ではない文書対応から、高精度な対訳文を抽出可能である。Masao Utiyama and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. In proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 72--79.

10 https://www.nict.go.jp/press/2014/07/28-1.html

11 また、国立研究開発法人産業技術総合研究所との連携で大規模データからの学習を実現した。https://www.aist.go.jp/aist_j/news/announce/au20170526.html

12 https://www.toshiba-sol.co.jp/news/detail/20190531.htm

13 http://www.soumu.go.jp/main_content/000285578.pdf

14 https://www5.cao.go.jp/keizai-shimon/kaigi/cabinet/2019/2019_basicpolicies_ja.pdf の P46~47 に記載。

15 https://www.kantei.go.jp/jp/singi/tougou-innovation/

16 http://www.moj.go.jp/hisho/seisakuhyouka/hisho04_00066.html

17 http://h-bank.nict.go.jp/index.html

している。これによって、分野毎に良質なデータを大量に収集することが可能となった。

分野毎に対訳データをNICTに集め、これらをまとめて分野毎に高精度機械翻訳をNICTが実現し広く利用可能とするという仕組である（図1）。

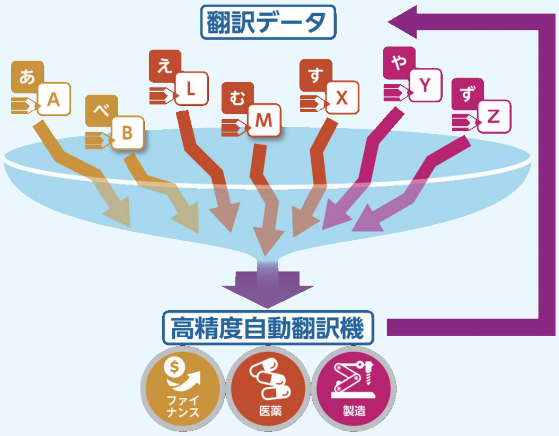

図1　翻訳バンクの概念

実際に、アストラゼネカが製薬の対訳データを寄付して高精度翻訳を実稼働に持ち込み、トヨタが同様に自動車法規のデータを寄付して外国語法規の高精度翻訳を実稼働している。また、他の複数の分野での試行も始まっている。

このように多分野で翻訳バンクが順調に進み始めたので、多数の分野での高精度化の数年内実現も視野に入りつつある。

翻訳メモリ、テキスト、Word、Excelなどの形式の日本語と他言語、英語と他言語の翻訳データの提供を募っている。Webサイト「みんなの自動翻訳」には、対訳集の登録を受け付けるページが用意されている。また、NICTと二者間契約を締結することも可能である。

機密保持契約や著作権などに対する不安の声を聞くこともあるが、公知の翻訳テキストを翻訳バンクへ提供することには法的な問題は全くない[18]。

7 AI翻訳の今後

AI翻訳の特徴的な誤訳に、原文の情報の一部が訳出されない（訳抜け）や原文に無い情報が訳文に表れる（湧き出し）があり、訳文が流暢であることから、これらに誤訳を見逃しやすいことが深刻な欠点と言われている。しかしながら、世界中の研究者の努力でこれも頻度が低くなりつつある。

また、現在のAI翻訳は文毎に翻訳する仕組みで文脈を考慮していない。そのために、同じ単語の訳が文毎に違うという現象が頻繁に起こり、特に、産業翻訳での活用時に忌避要素となっているが、1、2年で解決できるとみる研究者が多い。

このほか、逐次通訳から同時通訳への発展やマルチモーダル情報を参照することによって、文の翻訳の誤訳を解消する試みもある。

AI翻訳は、改良の速度が速く、研究開発は活況を呈しており、成長が期待できる、とても楽しみな存在となっている。

8 おわりに

本稿では、AI翻訳の到達点が、①翻訳精度が十分高く役に立つこと、②特に、特許文献の場合、特許庁がAI翻訳を一般公開しており広く活用されていること[19]、③AI翻訳は発展途上であり、翻訳精度がまだプラトーに到達してないこと、④文脈情報の活用も研究されており、訳語の統一等と遠くない将来に実現されそうなこと等を報告した。

また、⑤官の機械翻訳の研究・開発・普及への取組、⑥AI翻訳の可能性を最大化する対訳データ収集の活動として創出され、軌道に乗った翻訳バンクについて説明した。

18 http://h-bank.nict.go.jp/seminars/download/20190306/taichikakimuma190306.pdf

19 特許庁のサイトの評判はよく、特許庁への提出書類の翻訳等での有用性に対する異論はほぼないし、出願時の活用についてもNIPTAの「ユーザーから見たNMTの使い勝手と活用の展望」(http://aamtjapio.com/symposium181207.html)等前向きの取組が活発に行われている。