機械翻訳の原理と研究動向

呉　浩東

Research Trends and the Principle of Machine Translation

GO Kotoh

This paper addresses the mechanism of machine translation and its new trends in development. Firstly, we describe the traditional machine translation such as rule-based machine translation that transfers source language to target language using hand-crafted knowledge from lexical level, syntactic level and semantic level. We then introduce the mechanism of example-based machine translation. We then present the statistical machine translation based on bilingual corpus and machine readable dictionary. In recent several years, researchers develop neural machine translation that employs neural network to transfer languages with very high accurate rates. Finally, we in this paper discuss the problems of neural machine translation model and its possible solutions, and propose the future directions for machine translation development.

１．はじめに

近年、グローバル化が急速に進み、多言語への翻訳、しかもタイムリーでコスト対パフォーマンスのよい翻訳のニーズが高まる中、どうしたら品質を保ちつつ短時間でコストを抑えた翻訳ができるかということが大きな課題となってきた。

翻訳とは、「ある言語のテキストを別の言語の等価なテキストに置き換えること」である。変換元の言語は原言語（source language）、変換先の言語は目的言語（target language）と呼ばれる。機械翻訳ではこの翻訳タスクをコンピュータによって実現するものである。翻訳の性質から見ると、以下のように分類できる。

１．単語と構造の等価性を重視する翻訳
２．意味内容の等価性を重視する翻訳
３．効果の等価性を重視する翻訳

機械翻訳（Machine Translation, MT）は1947年にWarren Weaverによって提起された概念で、すでに60年余りの歳月が経っている。近年、コンピュータの容量と速度が急速に向上し、機械翻訳は飛躍的に成長を成し遂げている。また、国際化が進み、機械翻訳のニーズが非常に大きくなっている。一例として、Google社だけでも毎日の機械翻訳サービスの量は1,000億語を超える。

２．機械翻訳の難しさ

機械翻訳を実現するために、語彙的なずれや多義性、省略の問題を対処しなければならない。例えば、日本語に、the, a, anが存在しない。また、日本語では、主語の省略が多い。英語では主語が必要である。「○○さん」の場合、英語のMr.あるいはMs.に翻訳するかが不明である。また、日本語では単数・複数を無視することが珍しくない。英語には、助数詞（一双、二台、三着など）がない。さらに厄介な問題は言語における曖昧性問題である。語義の多義性について、runという動詞一つ取っても、run a mile〔１マイルを走る〕、run a test〔テストを実施する〕、run a store〔店を運営する〕ではまったく意味が違ってくる。そのほかに、「連続得点」「追い詰める」「伝線」など特定の文脈でしか利用しないケースも多い。**表 1** はさまざまな文脈における「掛ける」を英語に訳す例を示す。

語彙レベルで、敬語や謙譲語などの丁寧さの対応、固有名詞の判別、省略された単語の復元（「花子にあった」→「I met Hanako」）、文章全体的訳語の一貫性など、機械翻訳の実現には多くの難しい問題が伴う。

原言語と目的言語文内の単語の並び替え（アライメント）は機械翻訳におけるもう一つの難問である。文として文の正しい並び替えを実現するために、文全体の文法構造や意味を考慮しなければならない。

これは日本語と英語のような文法が大きく異なる言語では大変な対応策を講じる必要がある。例えば、「来週妹が飛行機で沖縄に行く」（**図 1**）の中に英語に訳すために語順変換が必要である。

また、日本における省照応問題の解決も機械翻訳にとっても難問として知られる。たとえば、日本語では指示代名詞（「あれ」「これ」など）が頻繁に使用

されるが、英語はこれらを明確に文に含める必要がある。そのため、日英翻訳では指示代名詞の指示先を特定する必要がある。しかし、複雑な文章に対して簡単には特定できないケースが多い。

文例	英訳
迷惑を掛ける	cause an inconvenience
鉄道を掛ける	build a railway
コートを掛ける	Hang a coat on
お金を掛ける	spend money
エンジンを掛ける	start the engine
腰を掛ける	sit down
水を掛ける	pour the water
布団を掛ける	put a quilt
目覚ましを掛ける	set the alarm clock
眼鏡を掛ける	Wear glass
DVDを掛ける	play a DVD

表１.「掛ける」を英訳する際の多様性

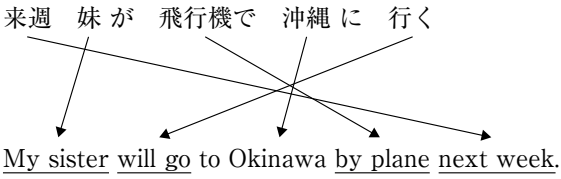

図１. 日英対訳文の並び替え例

３. 古典的な機械翻訳

機械翻訳には、解析を深くすれば意味の理解度が向上し、最終的に原言語と目的言語の意味は一致するという考え方と、言語はあくまで個別的なもので、解析を深くしても両者の意味は一致しないとする考え方がある。前者を代表する翻訳方式は**中間言語方式**、後者を代表する翻訳方式は**トランスファー方式**と呼ばれる。**図２**は中間言語方式とトランスファー方式の使い分けの可能性を提示するものである。

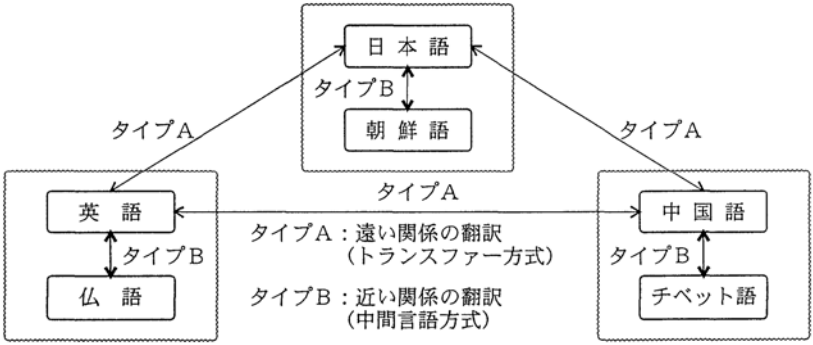

図２．語系と古典的機械翻訳

トランスファー方式では、原言語の文の単語レベル、統語レベル、意味レベルの処理を行い、得られた単語対応、統語（構文）構造、意味構造を目的言語の対応する構造に変換し、その構造に基づいて目的言語の文を生成する。単語レベルで変換を行う場合、**単語直接方式**と呼ばれる。統語構造間で変換を行う場合、**統語トランスファー方式**と呼ばれる。意味構造で変換を行う場合、**意味トランスファー方式**と呼ばれる。この３種類のトランスファー方式は**図３**を参照する。

中間言語（英：Pivot language）は、任意の言語を異なる任意の言語へ翻訳する際に利用する中間的な人工言語もしくは自然言語である。これに介した翻訳は**中間言語方式**あるいは**ビボット（pivot）方式**と呼ばれる。しかし、対象分野を限定しない中間言語の設計は極めて難しいため、ビボット（pivot）方式の使用は限定されている。また、分野を限定させると、世界知識の利用は可能であるため、**知識ベースに基づく機械翻訳**と呼ばれることもある。知識ベースに基づく機械翻訳では知識表現の記述力や効率的な操作に重点を置き、対象分野の知識を概念体系化することで意味的に深いレベルでの翻訳を目指している。中間言語に相当する意味表現と対象分野の概念知識を操作性の高い共通の枠組みで記述する。

古典的機械翻訳の開発の人手であるため、言語知識と世界知識のルールの作成作業のコースとが極めて高い分野を限定しないと翻訳精度が低く利用しづらい状況である。

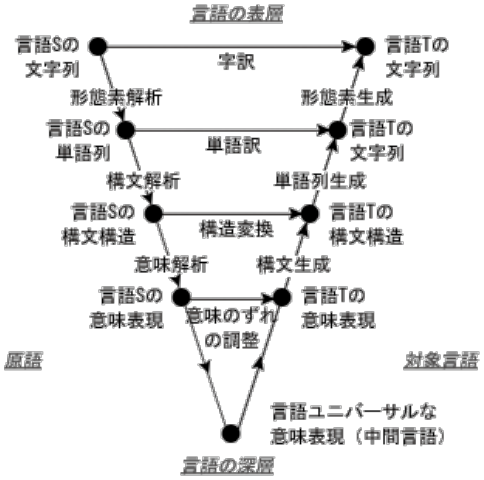

図３．トランスファー機械翻訳のイメージ

４．用例に基づく機械翻訳

用例に基づく機械翻訳（Example Based Machine Translation, EBMT）は、翻訳規則ではなく多量の翻訳用例を用意し、それを利用する翻訳方式である（長尾［１］）。用例ベース翻訳（EBMT）翻訳用例の獲得、辞書やシソーラス（類語辞典）などの言語資源を積極的に利用し、アドホックなルールを利用することが多い。例えば、

用例：私は学校で雑誌を読んだ！　→　I read a magazine in the school.
入力：私は家で新聞を読んだ！
出力：I read a newspaper in home.

EBMTでは、翻訳規則の代わりに用例集合（対訳コーパス）を用いることにはいくつかの利点がある。まず、用例は独立性が高い、翻訳規則で必要になる適用条件や相互関係を明示する必要がない。EBMTの仕組みは**図 4** に示される。

例えば、「名詞句の名詞句」における「の」を訳す方法は**表 2** のようにすれば、複雑な条件を規則として記述する必要がない。さらに、用例の追加は翻訳規則（翻訳ルール）の追加に比べて衝突や副作用が少ないため、それに応じた翻訳の質の向上が期待できる。

AのB	英語訳
八日の午後	the afternoon of the 8th
会議の参加費	the application fee for the conference
三つのホテル	three hotels
京都のホテル	hotels in Kyoto
太郎の母親	Taro's mother
崖の上	on the cliff

表２．「名詞句の名詞句」の意味構造の例

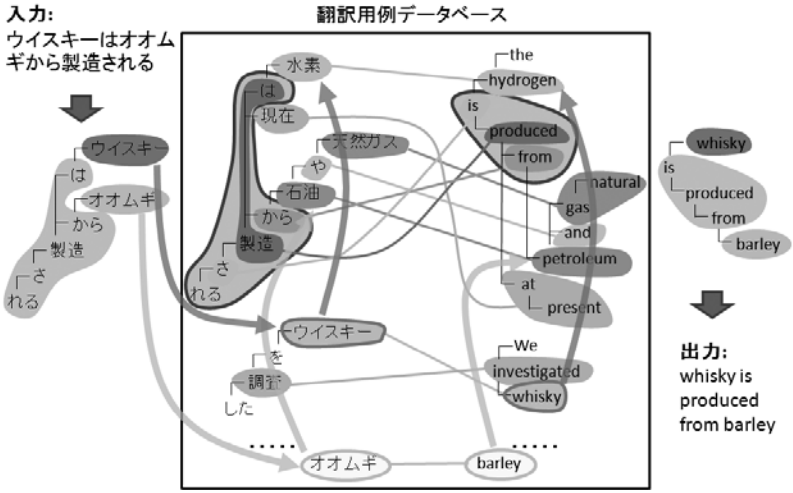

図４．EBMTの仕組み

出典：黒橋禎夫　研究紹介
（http://nlp.ist.i.kyotou.ac.jp/index.php?%E7%A0%94%E7%A9%B6%E7%B4%B9%E4%BB%8B）

５．統計的機械翻訳

統計的機械翻訳（Statistical Machine translation, SMT）は1990年頃に提案された翻訳方式である。さまざまなコーパスなどから確率的パラメータを学習し、基本的に辞書などの言語資源は利用しない、頑健な数学的知識に基づいているといった特性を持っている。**図５**は**統計的機械翻訳**の原理を示す。

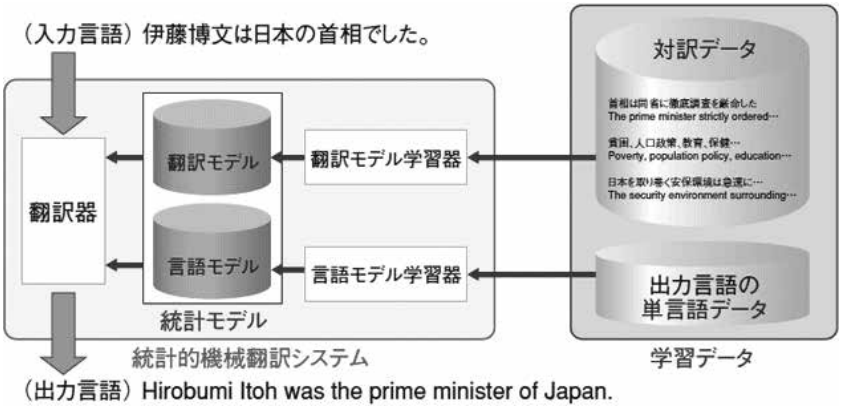

図５．統計的機械翻訳の仕組み

出典：Statistical machine translation, hereafter referred to as SMT（NTT）

統計的機械翻訳は大別すると、IBMモデルと句（フレーズ）に基づく統計翻訳モデルがある。早期の手法は、単語を単位として翻訳するモデルを用いており、文脈の情報を活用しにくいという課題があった。そこで、フレーズベース機械翻訳が提案された。当方法はシンプルで、いままでに述べた手法に比べて容易で、翻訳のスピードも速いことが人気を集めた理由であった。ここで、フレーズベース機械翻訳について説明する。フレーズベース機械翻訳では、１）翻訳は原言語文を句への分割。２）句単位で翻訳。３）翻訳された句を並び替え、目的言語文を生成する。

フレーズベース機械翻訳はまず原言語と目的言語の句対応を抽出する。ここでは、英語から日本語への翻訳を考えると、図６に示すように、句の並び替えを行う。さらに、句翻訳の確率は、次のように対訳コーパスの中の対応関係から最尤推定で求めることができる。そのなかで、句の対応付け（alignment）をａで表現し、文ｅのｉ番目の語が文ｊのai番目の語に対応付けられるとする（**式１**）。対訳文ｅ、ｊはさまざまな対応付けを介して得ることができるので、**翻訳モデル**の仕組みは**図６**に示される。

$$P(e \mid j) = \Sigma_a P(e, a \mid j) \quad (1)$$

なお、ここで得られる翻訳システムは、英日翻訳を行うものである。

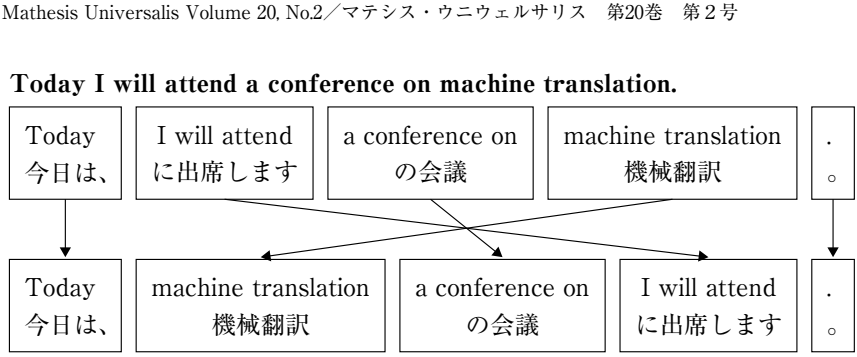

図６．フレーズベース機械翻訳における句の並び替え

図７はフレーズベース機械翻訳の全体像を示す。大量の対訳コーパスを登録し統計的手法により訳文を生成する方法である。ルールや辞書の開発の必要がなく、原文と訳文の両方の言語の性質に縛られにくいため、多言語化が容易だとされている。

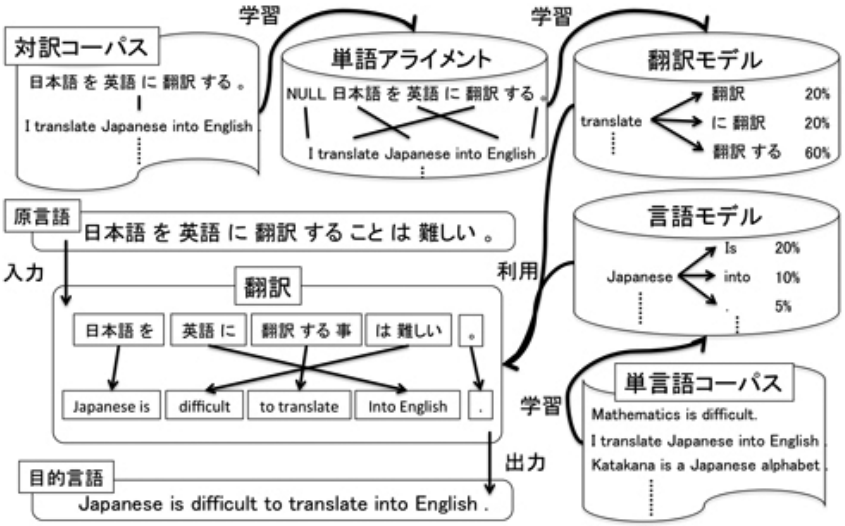

図７．フレーズベース機械翻訳の仕組み

出典：http://www.ipsj.or.jp/magazine/hakase/2016/NL01.html

６．ニューラル機械翻訳

６．１　ニューラル機械翻訳の特徴

ニューラル機械翻訳（Neural Machine Translation, NMT）は、１つの脳をモデルにしたマルチ処理デバイスである大規模な**ニューラルネットワーク**を通じて機械に翻訳を学習させる新たなアプローチである。NMTは、素性を自動的に学習するニューラルネットワークにより２言語の対応付けを数値ベクトルで表現し、エンコーダー・デコーダーモデル（encoder-decoder model）でMTを実現する。原言語の文をエンコーダーによって数値ベクトルで表現し、デコーダーで逐次可変長の目的言語の出力シンボルを生成する。

SMTでは、入力文を単語列に分割し、それぞれの単語列を翻訳してつなぎ合わせることで訳文を生成する。これに対してNMTでは、まず入力文の各単語を分散表現と呼ばれる数値の並びに変換し、これらを合成して入力文全体を表す「文の分散表現」を得る。そして単語の入力が終わると、この文の分散表現に対応する訳語リストのスコアを計算し、最もスコアの高い訳語を出力する処理を繰り返す。

NMTは、機械翻訳研究者や開発者の間で人気が高まっている。学習したNMTシステムは、多くの言語ペアにおいて、語句に基づいた統計ベースの翻訳よりも優れた翻訳パフォーマンスを見せ始めている。

６．２　ニューラル機械翻訳の仕組み

NMTはディープラーニングを採用した機械翻訳の方式である。**図８**はNMTの仕組みを示す。対訳データを使って原言語の学習を行い、学習を完了することで入力した文章を目的言語に翻訳する。その流れは以下のように展開する。

１）原言語の文章の形態素解析を行い、単語列に分割する。
２）分割した単語を数値表現に変換する。
３）単語の符号化を行い、ベクトルに変換する。
４）エンコーダー（encoder）は原言語の文の符号化をし、原言語の文章のベクトルを作成する。この層はいままで作成したベクトル群を再帰ニューラルネットワークによって処理する。ここでは、符号化された入力文のどこに注目すべくアテンション機構（attention mechanism）を用意する。
５）Step4で作成された原言語のベクトルとアテンション情報をもとに、デコーダー（decoder）を用いて目的言語のベクトルに変換し、目的言語の単語を順次に生成し、新しいベクトルを生成する。

６）Step5の同じ処理を繰り返して実行し、文章の終わりを示す特殊文字EOS（End Of Sentence）の出力を完成後、翻訳作業を終了する。

NMTにおいては、エンコーダー、デコーダーとアテンション機構は重要な役割を果たしている。簡単に言えば、エンコーダーでは、まず各単語を分散表現と呼ばれる数百次元からなる実数値ベクトルに変換する。そこで次に各単語の先頭からと末尾から１つずつに読み込み、RNN（Recurrent Neural Network）によって前または後ろの単語を考慮してベクトル表現を作り出す。アテンション機構は、エンコーダーで処理された文、次の単語を訳出する際に注目すべき箇所を判断する。デコーダーは１つのRNNで構成されており、コンテキストを反映したベクトルを１つ前の単語の情報を受け取り、次の単語を出力する（中澤［25］）。

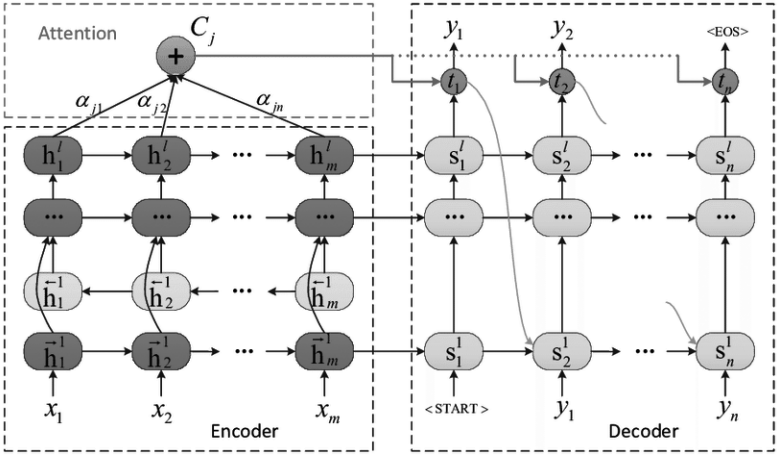

図８．NMTの仕組み

出典：Loug Zhou, etc. Look-ahead Attention for Generation In Neural Machine Translation. 2017

６．３　問題とチャレンジ

Encoder-to-Decoderの構造に基づくNMTは通用性があるモデルである。もともとは機械翻訳のために設計されたものではない。そのため、以下のような問題が生じる。

1）未登録語の対応問題とその改善策。
翻訳処理時間と空間におけるコストをコントロールするために、NMTは原言語と目的言語の間に5万語前後規模の語彙表を採用する場合が多い。カバーできていない語彙の影響で、原言語の文の語彙情報が不足という問題を生じる一方、出力された文の理解度（翻訳の忠実さ）に影響も出る（Jean ら［1］，Arthur ら［29］）。
2）先行知識の有効活用。
3）アテンション機構（attention mechanism）のメカニズムのさらなる改善。現在のNMTは、約束が不備のため、訳抜けや訳重複（約4％）の問題が生じる（中澤［25］）。
4）ニューラルネットワークの改良。

近年、NMTの研究開発に大きな関心が集まっている。語彙表規模の改善（Jean ら［2］，Luong ら［3］）、アテンション機構の改良（Tu ら［7］，Cohn ら［5］）、NMTとSMTの結合（He ら［6］，Stahlberg ら［8］）、言語知識の導入（Eriguchi ら［9］，Sennrich ら［10］，勝俣 ら［31］）、単言語コーパスの使用（Gulcehre ら［11］，Spennrich ら［12］，Cheng ら［13］，Zhang ら［14］）、記憶メモリの使用（Wang ら［15］）、NMTモデルの訓練（Wuら［17］）などの研究が挙げられる。

6.4　ニューラル機械翻訳についての考察

NMTの特長には以下のようなものがある。
1）人工的なニューラルネットワークが自律的に学習するアルゴリズムを採用している。
2）フレーズにとどまらず、文章全体を考慮する。
3）言語の持つニュアンス、語尾変化や敬語、男性／女性用語を学習する。

これらの特長を有することにより、統計的機械翻訳と比べて、語順、構文エラーといった問題が発生しにくく、また、韓国語、日本語、アラビア語といった文法と語彙が難解だとされる言語にも適切に対応できるとされている。

翻訳の流暢さと正解度において、NMTはSMTより優れている。また、SMTにとって困難である複雑な単語と長距離の対応付けにも、NMTはうまく対応できる。一方、翻訳の忠実さにおいて、NMTはまた改良すべき点が多く残っている。

2016年に、GoogleのGNMT（Google's Neural Machine Translation）が登場した。大規模対訳コーパス、巨大なNMTモデル、大量のGPUを生かして高精度な機械翻訳を実現している。GNMTの翻訳例は**図９**にまとめ、そのパフォーマンスは**図10**に示す。

ニューラル機械翻訳（NMT）は、１つの脳をモデルにしたマルチ処理デバイスである大規模なニューラルネットワークを通じて機械に翻訳を学習させる新たなアプローチです。
Neural machine translation（NMT）is a new approach to let machines learn translation through a large-scale neural network which is a multi-processing device modeled on one brain.
神经机器翻译（NMT）**是一种让机器的新方法通过大规模的神经网络学习翻译处理设备模仿一个大脑**。

図９．Googleによる日英中機械翻訳の翻訳例

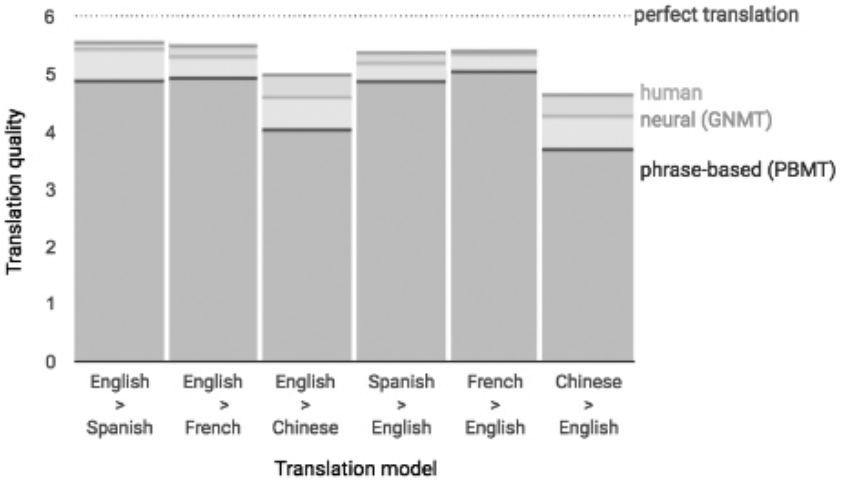

図10．Googleのニューラル機械翻訳の精度

出典：Google Research Blog

NMTの質は翻訳の精度がこれまでの他手法より飛躍的に向上していることを体験できる。**図10**はGoogleが発表している2016年時点のGoogle翻訳に関するグラフである。単語対によって人間の翻訳にほぼ同じ程度の質を有する

（Google［29］）。2018年に、Microsoftの研究者らは、中国語のニュース記事の文章を人間と同じ正確さで英語に翻訳する世界初の機械翻訳システムを開発したと発表した（Microsoft［2018］。https://japan.cnet.com/article/35116178/.）。

7．まとめ

ニューラル機械翻訳は多くの言語対で実用的なレベルまで達成している。特に、最近はニューラル機械翻訳を実現することにより、英仏、英西のような近い言語対で人間に近いパフォーマンスを発揮している。ウェブなどから対訳データを大量に収集できる分野、大規模の対訳データ（対訳コーパス）や、対訳の言語資源が豊富な言語間は高精度な翻訳が実現している。その反面、大きく異なる言語対では単語対応が取りにくくローカルな情報で解決できない問題の対応は大変難しい。大規模な学習コーパスが存在しない場合や、少数言語や対訳データが存在しない分野、独自性の高い言語、書き方（口語）、深い意味理解と文脈理解を必要とする文学作品の機械翻訳には数多く難問は残されている。これからは、各翻訳方式の特徴を生かし、特に、SMTとNMTでは同じ文でも翻訳結果が異なることが多く、両手法を補完する方法が考えられる。今後、さらなる性能の高い翻訳方式の改良も今後の課題として力を入れるべきと思われる。

参考文献

［1］Nagao, Makoto, A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, Artificial and Human Intelligence, 1984.
［2］Jean S,, Cho K., Memisevic et al. On Using Very large target Vocabulary for Neural Machine Translation in Proc, of ACL, 2015.
［3］Luong M. t., Sutskever I., Le Q. V., et al. Addressing the Rare Word Problem in Nueral Machine Translation. in Proc, of NAACL, 2015.
［4］Long Zhou, Jiajun Zhang, Chengqing Zong. Look-ahead Attention for Generation in Neural Machine［Cs. CL］, 2017.
［5］Cohn T., Cong D. V. H.. et al. Incorprorating Stuctural Alignment Biases into an Attentitional Neural Translation. in Proc, of NAACL, 2016.
［6］He W., He Z., Wu H. et al. Improved Neural Machine translation with SMT Features. in Proc, of AAAI, 2016.
［7］Tu Z. Lu.,Liu Y., et al. Modeling Coverage for Neaural Machine Translation, in Proc, of ACL 2016.
［8］Stahlberg F., Hasler E., Waite A., et al. Syantactically Guided Neural Machine Translation. in Proc, of ACL, 2016.

［９］Eriguchi A., Hashimoto K., Tsuruoka Y.. Tree-to-Attentional Neaual Machine Translation. In Proc. Of ACL, 2016.
［10］Sennrich R., Haddow B.. Linguistic Input Features Impove Neural Machine Translation. in Proc, of First Conf. on machine Translation, 2016.
［11］Gulcehre C.Firat O. et al. On Using Monolingual Corpora in Neral Machine Translation. J. of Computer science, 2015.
［12］Sennrich R., Haddow B., Birch A.. Improving neaural machine tranlation Models with Monolingual Data. in Proc, of ACL, 2016.
［13］ChengY., Wu. W., He Z.. et al.. Semi-Supervised Learning for Nueral Machine Translation. in Proc, of ACL, 2016.
［14］Zhang J., Zong C.. Exploiting Source-side Monolinual dta in Neaural Machine Translation. in Proc, of EMNLP, 2016.
［15］Wang M. T., Le Q. V.. et al.. Multi-task Sequence to Sequence Learning. in Proc, of ICLR, 2016.
［16］M., Lu Z., et al.. Memory-enhanced Decoder for Neural Machine Translation. . in Proc, of EMNLP, 2016.
［17］Wu Y., Schuster M., et al.. Google's Neaural Machine Translation System: Bridging the Gap between human and Machine Translation. J. of arvix, 2016.
［18］三木光範，加藤恒昭．自然言語処理，共立出版，2014
［19］黒橋禎夫，自然言語処理，放送大学教育振興会，2015
［20］小町　守　監修，奥野　陽，グラム・ニュービッグ，荻原正人，自然言語処理の基本と技術，翔泳社，2016.
［21］長尾　真（編）．自然言語処理，岩波書店，1996.
［22］長尾　真，黒橋禎夫，佐藤理史，池原　悟，中野　洋．言語情報処理，岩波書店，1997.
［23］奥村　学．自然言語処理の基礎，コロナ社，2005.
［24］吉田　仙，水島昌英，田中公人．ニューラル機械翻訳によるサービス創造に向けた取り組み，NTT技術ジャーナル，pp.34-37，2018.
［25］中澤敏明，機械翻訳の新しいパラダイム　ニューラル機械翻訳の原理，情報管理，vol.60，no.5，pp.299-306，2017.
［26］根石将人 ら，ニューラル機械翻訳における埋め込み層の教師なし事前学習，情報処理学会研究報告，vol.2017-233．no.1，2017.
［27］後藤功雄．機械翻訳技術の研究と動向，NHK技研　R&D/No.168，2018.
https://qita.com/kenrohmiyoshi/items/8d767242da8ec87b8962.
［28］**李亚超，熊德意，张民．神经机器翻译综述．计算机学报**，online，2017.
［29］“A Neural Network for Machine Translation, at Production Scale”, Google research blog.
https://research.googleblog.com/2016/09/q-neural-network-for-machine-translation.html.
［30］Arthur P., Neubig G., Nakamura S. Incorporating discrete translation Lexicons into Neural Machine Translation. In Proc. of EMNLP, 2016.

［31］勝又　智，松村雪桜，山岸駿秀，小町　守．ニューラル機械翻訳における共起情報を考慮した語彙選択，言語処理大会第24回年次大会発表論文集，2018.
［32］奥村　学，渡辺太郎，今村賢治，賀沢秀人，Graham Neubig．機械翻訳（自然言語処理シリーズ）コロナ社，2014.