統計方式機械翻訳とニューラル方式機械翻訳のハイブリッドシステム

Hybrid System of Neural Machine Translation and Statistical Machine Translation

元山梨英和大学教授
江原　暉将

1967 年早稲田大学理工学部卒。同年 NHK 入局。2003 年 諏訪東京理科大学教授。2009 年 山梨英和大学教授。2015 年 退職。アジア太平洋機械翻訳協会（AAMT）／Japio 特許翻訳研究会委員。

1　はじめに

機械翻訳は、規則方式、用例方式、統計方式、ニューラル方式、と進歩を重ね、今日に至っている。まず、これらの諸方式を概観してみよう。

規則方式は、機械翻訳の研究が始まった当初から行われており、人が構築した対訳辞書と文法規則を用いて翻訳を進めるものである。ちょうど我々が英語を習い始めのときに辞書と文法を使って英文和訳を行うときの状況に似ている。翻訳対象の英文の各単語を英和辞書を使って調べ、どのような日本語単語に訳せばよいかを知る。そして英語のどの部分が主語であり、動詞であり、目的語、前置詞句であるかを調べ、それらを日本語の文法に従って並べ替えて和訳を得る。日本語話者であれば日本語の文法は知っているので和訳を生成するのは容易であるが、コンピュータには日本語を生成する日本語文法も教えておく必要がある。

このような規則方式で、一応の翻訳はできるが、言語表現には曖昧性があり、それを解消しないと良い翻訳は得られない。有名な例で “Time flies like an arrow.” という文がある。この文は、「時は矢のように飛ぶ。」（光陰矢のごとし）と訳すのが普通であるが、「時蝿は矢を好む。」という解釈も可能である。このように複数ある解釈の中から、コンピュータが適切な訳文を選択するには、どのような知識が必要であろうか。その知識を過去の翻訳用例から学ぼうというのが用例方式である。まず過去の翻訳用例をたくさん集め、翻訳用例データベースを構築する。そして翻訳対象の原文を翻訳用例データベースの中で検索し、マッチした用例の訳文を採用する。原文全体ではなく、一部分がマッチする用例を組み合わせて訳文を組み立てることも行われる。これが用例方式である。このような翻訳用例の利用は、その後の統計方式、ニューラル方式にも引き継がれている。

統計方式は用例方式の発展形であり、翻訳用例を統計的な標本とみなし数学モデルを強化した方式である。その結果、統計学や機械学習の手法が応用でき翻訳精度も向上した。典型的な統計方式である句レベルの統計方式翻訳では、phrase table と reordering table という二つのテーブルに翻訳知識が集約されている。前者は規則方式での対訳辞書に相当し、後者は文法規則（語順を入れ替えるという単純な文法ではあるが）に相当する。

ニューラル方式は単語を数値ベクトルで表すことから始まった。従来の機械翻訳は、単語を単なるテキストとして扱っていた。例えば「コンピュータ」と「計算機」は異なる単語であり、両者の意味的類似性は考慮できなかった。しかし、単語をベクトルとして扱うことで、意味的類似性をベクトル間の近さで表すことが可能になった。さらに文の構文構造や文脈までベクトルで表し、ベクトル間の処理を人の神経細胞を模したニューラルネットワークで行うのがニューラル方式機械翻訳である。ニューラル方式は期待以上の翻訳精度を実現し、現在、機械翻訳の主流となっている。

しかしニューラル翻訳であっても完全ではなく、課題がある。典型的な課題として、不足翻訳（under translation）と過剰翻訳（over translation）の問題がある。不足翻訳は翻訳すべき重要な情報が抜けてしま

うことで訳抜けとも呼ばれる。逆に過剰翻訳は不要な情報が（しばしば繰り返して）訳出されることで湧き出しとも呼ばれる[1]。これらはニューラル方式翻訳でしばしばみられ、課題の一つとなっている。不足翻訳や過剰翻訳の実例は３節を参照されたい。

規則方式から始まって用例方式、統計方式、ニューラル方式と見てきたが、それらの各方式には、それぞれ長所と短所があり、各種方式を組み合わせることで長所を生かすことが可能である。そのような例として統計方式とニューラル方式を組み合わせた方式を紹介しよう。

2 統計方式とニューラル方式の組み合わせ

前節で説明したようにニューラル方式の欠点の一つとして不足翻訳や過剰翻訳がある。一方、統計方式にはそのような誤訳は比較的に少ない。そこで両者を組み合わせてより良い機械翻訳を実現することが考えられる。総合的にはニューラル方式の方が統計方式よりも翻訳精度が高い。そこで、ニューラル方式を基本として、その中で統計方式を利用するハイブリッド方式を考える[1]。

ハイブリッド方式システムの構成を【図 1】に示す。翻訳対象の入力文はニューラル方式翻訳機（NMT）と統計方式翻訳機（SMT）の両方で処理される。SMT は翻訳結果を一つ出し、NMT は翻訳結果を複数（【図 1】では n 個）出力する。n 個の NMT 出力はシステムが曖昧性を解消しきれなかった範囲から優先度の順に出力するものである。これら n 個の NMT 出力は SMT 出力と比較され、SMT 出力に最も近い NMT 出力を選択して（【図 1】では NMT 出力 i を選択）、システム翻訳結果として出力する。NMT 出力と SMT 出力の近さを測るメジャーとしては機械翻訳の自動評価基準を用いることができる。文献[1]では IMPACT[2] を用いている。SMT 出力には不足翻訳や過剰翻訳が比較的に少ないので、それに近い NMT 出力にも不足翻訳や過剰翻訳が少なくなることが期待できる。

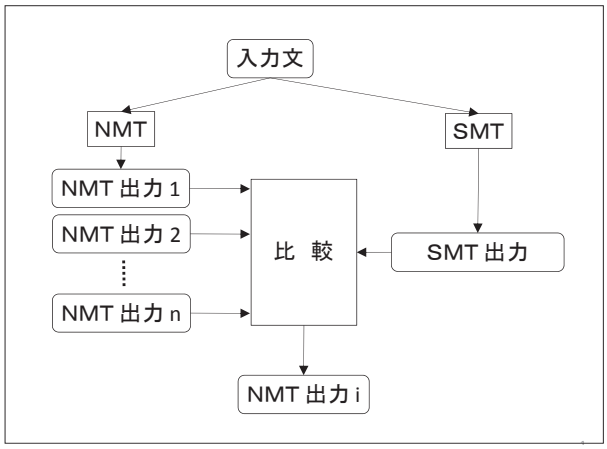

図 1　ハイブリッド方式システムの構成図

3 翻訳例

英日特許翻訳から３つの翻訳例を選び【表 1】に示す[1]。表中、原文は入力の英文、基準訳文は参考のために示した日本文、ハイブリッド出力は前節で説明したハイブリッドシステムの訳文、NMT 出力は NMT の訳文（【図 1】では NMT 出力 1 に相当）、SMT 出力は SMT の訳文である。

例文 1 では、NMT 出力に極端な不足翻訳がある。SMT 出力には不足翻訳がないため、ハイブリッド出力にも不足翻訳が生じていない。

例文 2 では、NMT 出力に不足翻訳と過剰翻訳が両方生じている[2]。これも SMT 出力を利用することでハイブリッド出力では修正されている。

例文 3 では、NMT 出力に過剰翻訳があるが、ハイブリッド出力にはない。

1　言語の違いを乗り越えて情報を伝達する翻訳では、原言語話者には不要だが目標言語話者には必要な背景情報を追加したり、逆に原言語話者には必要だが目標言語話者には必ずしも必要でない詳細情報を削除したりすることが良く行われる。これらは過剰翻訳、不足翻訳とは異なるものであり、筆者は、これらを追加翻訳（additive translation）、削除翻訳（subtractive translation）と呼んでいる。

2　「コマンド（ACT、READ、」以下文末までが不足しており、「図 4 の回路のシミュレーション結果の一部」が繰り返しており過剰である。

表 1　ハイブリッド方式と他の方式の出力比較例

1	原文	The oldest is the capacitive divider bridge, one arm of which consists of a reference capacitance and the other of the capacitance to be measured.
	基準訳文	最も 古い もの は 容量 分圧 器 ブリッジ であり 、 その 1 つ の アーム は 基準 静 電 容量 から 成り 、 その他 の 静 電 容量 が 測定 さ れる 。
	ハイブリッド出力	最も 古い もの は 容量 分圧 ブリッジ であり 、 その 1 つ の アーム は 、 測定 さ れる べき 基準 キャパシタンス と 他の キャパシタンス と から 成る 。
	NMT 出力	最も 古い 。
	SMT 出力	最も古いものは、容量性ディバイダブリッジ、基準キャパシタンスと測定されるべきキャパシタンスの他の 1 つを構成するアームである。
2	原文	FIG. 5 shows a portion of waveforms of a simulation result of the circuit in FIG. 4, and commands (ACT, READ, PRE), GBSTB, GBSTT, FXb 2, PUMP, FXB, FXT, SWL, and RFX are shown as signals.
	基準訳文	図 5 に は 、 図 4 の 回路 の シミュレーション 結果 の 波形 が 一部 示さ れて おり 、 信号 と して 、 コマンド （ ACT 、 READ 、 PRE ） 、 GBSTB 、 GBSTT 、 FXb 2 、 PUMP 、 FXB 、 FXT 、 SWL 、 RFX が 示さ れて いる 。
	ハイブリッド出力	図 5 は 、 図 4 の 回路 の シミュレーション 結果 の 一部 を 示す もの であり 、 コマンド （ ACT 、 READ 、 PRE ） 、 GBSTB 、 GBSTT 、 FXb 2 、 PUMP 、 FXB 、 SWL 、 RFX を 信号 と して 示す 。
	NMT 出力	図 4 の 回路 の シミュレーション 結果 の 一部 と 、 図 4 の 回路 の シミュレーション 結果 の 一部 を 示す 図 である 。
	SMT 出力	図 5 は、図 4 の回路のシミュレーション結果の波形の一部を示す、及びコマンド（ACT、READ、PRE）、GBSTB、GBSTT、FXb 2、PUMP、FXB、FXT RFX SWL とは、信号として示されている。
3	原文	However, specifications sometimes differ from one image forming apparatus to another.
	基準訳文	しかし、画像形成装置ごとにスペックが異なる場合がある。
	ハイブリッド出力	しかし、ある仕様は、時々、1 つの画像形成装置とは異なる。
	NMT 出力	しかし、時々、1 つの画像形成装置とは異なる仕様が異なる。
	SMT 出力	しかし、仕様は、装置を形成する 1 つの画像から別の画像へ時には異なる。

4 まとめ

ニューラル方式機械翻訳は総じて良い翻訳を実現できるが、必要な情報が抜ける不足翻訳や不要な情報が湧き出す過剰翻訳という欠点がある。一方、統計方式機械翻訳にはそのような誤訳は比較的に少ない。そこで両者を合わせたハイブリッド方式を用いることでニューラル方式の精度の良い点を生かしつつ、不足翻訳や過剰翻訳を少なくすることができた。

これまで機械翻訳は、規則方式、用例方式、統計方式、ニューラル方式と発展してきた。しかしこれらの方式には一長一短があり、それぞれの長所を生かしたハイブリッド方式が有効である。特に実用機では、致命的な欠陥は許されず、それを回避する手段としてハイブリッド方式を採用することが考えられる。

参考文献

[1] Terumasa Ehara : SMT reranked NMT, *Proceedings of the 4th Workshop on Asian Translation*, pages 119–126, Nov., 2017.

[2] Hiroshi Echizen-ya, Kenji Araki : Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum, *Proceedings of the Eleventh Machine Translation Summit* (*MT SUMMIT XI*), Page. 151-158, Sept. 2007.