Googleは、2006年にSMTでスタートさせたGoogle翻訳サービスを2016年にNMTベースのGNMT (Google Neural Machine Translation [7,8]) に切り替え、精度を向上させ話題になった。このGNMTは、複数言語と英語間の翻訳用にトレーニングされたのち、わずかな改良により、非英語言語同士の翻訳に利用可能になった。

NMTでは、言語を数百次元の分散表現 (ベクトル化) とする「エンコーダー」、分散表現に対して、着目点を制御する「アテンション機構」、分散表現とアテンション機構を元に出力文を復号化する「デコーダー」から構成される [5]。エンコーダーにはフォワード/バックワード再帰型ニューラルネットワーク (forward/backward Recurrent Neural Network ; RNN) が接続され、前後の単語の情報 (コンテクスト) を考慮した分散表現をつくる。アテンション機構は、エンコードされた入力文とデコーダーの状態から次の単語を訳するときに注目すべき箇所を示すコンテクストベクトルを出力する。デコーダーは、RNNでコンテクストベクトルと一つ前に出力した単語の情報を受け取り、次の訳語を出力する。ただし、NMTでは、入力文の最後に置かれた「文末」記号をデコーダーが出力したときに翻訳が終了する。このため、翻訳終了が予測できず、翻訳過程も見えない。場合によっては、翻訳の重複や抜けが発生する。しかし、こうした問題があるものの、NMTは従来のSMTを上回る精度と応用範囲の広さを持っており、現在の機械翻訳研究の中心となっている。

言語処理技術の様々な応用

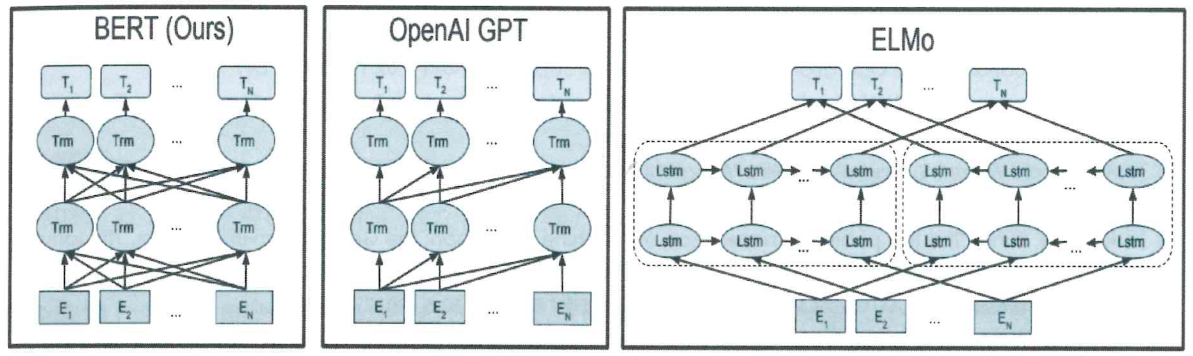
言語処理技術は機械翻訳だけでなく、言語に関わる各種認識処理などでも利用されることが多い。例えば、音声認識、音声合成、文字認識などで、単純な言語情報 (例えば単語辞書や、簡易な文法ルールなど) を利用して認識率などを向上させる試みは以前より行われていた。これらの処理では、音の文字のつながりの背後に確率的に遷移する状態を持つ確率モデルを想定している。このとき確率モデルは、隠れマルコフモデル (Hidden Markov Model ; HMM、ある時点の状態は、直前の状態にのみ依存するとしたマルコフ過程モデルにおいて、観測できない隠れた状態があるとしてつくられる確率状態遷移モデル) を前提にした機械学習が使われることが多かった。こうした統計的手法に対して、ニューラルネットワークを使う手法が登場している。音声合成 (Text-to-Speech) では、HMMを使った統計モデルを利用する場合、テキスト解析を行い、言語特徴量を求め、音響モデルを使って音響特徴量に変換し、最後に波形発生器を使って音声とするが、ディープニューラルネットワークを使う場合、言語特徴量から直接音響特徴量 (波形発生器への入力パラメーター) を生成する [9] ことや、言語特徴量から直接音声波形を生成する [10] ことなどが可能になる。

(3) 最新技術動向

Transformerによる自然言語処理能力の飛躍 (BERT、ELMo、GPT)

ニューラル自然言語処理で高い精度を達成するには、大量の学習データが必要であるが、様々なタスクのそれぞれについて大量の学習データを用意することは容易なことではない。そこでまず、質疑応答や自然言語推論といった様々な自然言語処理のタスクに共通的な汎用性の高いモデルを、大量のデータで事前トレーニング (Pre-training) しておき、それをベースにして個別のタスクごとの「ファインチューニング」(Fine-tuning) を行うというアプローチが検討されるようになった。ファインチューニングにおいては、個別のタスクごとに必要な学習データはそれほど大量でなくても良い。このような事前トレーニングの手法として特に注目されたのは、2018年に提案されたOpenAIのGPT[※37]、ELMo、GoogleのBERT[※38]である。これらは、RNNやCNNを使わずに、アテンション機構のみでベンチマークデータセットにおける質疑応答などのタスクにて従来よりも高い精度が達成可能なことを示したTransformerモデル[※39] (自己アテンション機構を使用) を活用している。OpenAIのGPTはleft-to-right単方向のTransformerモデル、BERTは双方向のTransformerモデル、ELMoはleft-to-rightとright-to-leftの連結モデルである。これらのモデルでは、事前トレーニングのアプローチに、文脈を考慮する仕組みも取り入れられている (図2-2-19)。

■図2-2-19 BERT、ELMo、GPTのニューラルネットワークの構造



出典：BERT、Pretraining of Deep Bidirectional Transformers for Language Understanding[※40]

自然な会話の実現 (Duplex)

スマートスピーカーの登場により、音声認識技術の一般への認知度は各段に上がるとともに、企業側も実用性を上げるための改良が行われてきた。Googleが2019年2月に発表したLive Transcribeは聴覚障碍者向けの書き起こし技術であるが、長年改良を続けてきたクラウド側のAutomatic Speech Recognition (ASR) をベースに、リアルタイム性、応答性など自然な情報提示を可能にしている。また2018年のGoogle I/Oで発表され話題となったDuplexでは、再帰型ニューラルネットワーク (RNN) を基盤に、電話会話データのコーパスを学習し、同社のTensorFlowベースの汎用機械学習プラットフォーム「TensorFlow Extended (TFX)」を使ってハイパーパラメーター最適化を行い、モデルを改善したとされる (図2-2-20)。自然な会話のために、「AlphaGo」で知られるDeepMindチームが開発した音声生成モデル「WaveNet」や、Google開発の音声合成技術「Tacotron」と波形接続型のテキスト読み上げエンジンを用い、状況に応じたイントネーションの制御をしているという。

※37 Radfordほか.Generative Pre-training Transformer,2018.
※38 Jacob Devlinほか.BERT : Pretraining of Deep Bidirectional Transformers for Language Understanding
※39 アテンション機構を単語間の相関を知って利用する仕組みとして利用したDNN。Googleが論文"Ashish Vaswani、Attention Is All You Need,2017"で発表。
※40 https://arxiv.org/pdf/1810.04805.pdf