第 6 節　英語技術文書の翻訳，要約への活用

(株) サイバーテック　橋元 賢次

はじめに

技術文書（Technical Document）とは，メーカーが販売および提供をする製品が要求事項を満たしていることを示す情報が含まれた文書であり，製品の適合性を立証するためのものである。主に必須となる要求事項をどのように適合させ，対策を実施したのかを記述し，製品がわかるような仕様書，もしくは配線図といった技術資料で構成されており，製品の製造終了日から 10 年間保管する必要がある。さらに，CE マーキングの自己宣言を行う場合は，技術文書を基に適合宣言を実施する。

したがって，技術文書は国内メーカーが提供する製品であるからといって，日本語だけで提供すれば良いという文書ではなく，日本語を原文とした上で英語を中心とした言語セットによる訳文も含めた形で構成されることが多い。つまり，日本語の技術文書を英訳するという事が求められるだけではなく，製品のローカライズを実施する場合は，英語以外の翻訳作業も必要となる。

また，技術文書は内容が高度であることが多く，ページ数も多くなりがちである。そのような技術文書から製品の情報を出来るだけ素早く理解するためには，要約された文章が存在すると製品に対する理解もスピーディに行う事が可能だが，要約作業は翻訳とはまた別の視点で行う必要がある。

1.　技術文書の翻訳手段

翻訳は一般的に「映像翻訳」「出版翻訳」「産業翻訳」の 3 種類に大別されるが，技術文書の翻訳は「産業翻訳」に該当する。また，翻訳手段として，従来は翻訳者による翻訳（人手翻訳）から始まり，その後 CAT（Computer-Aided Translation）と呼ばれる，コンピュータ支援翻訳ツールが登場することで，人手翻訳の効率化が向上したが，近年は機械翻訳や AI 翻訳の精度が一気に向上し，様々な産業翻訳作業において，機械翻訳や AI 翻訳は無くてはならないものになった。

2.　AI ブームのライフサイクル

従来の人手翻訳から，機械翻訳や AI 翻訳を語る上で，ベースとなる AI の発展過程に触れる必要がある。AI 技術を活用したシステムとしては，故障検出や画像生成など，様々なものがあるが，技術文書が関連する機械翻訳や AI 翻訳が発展してきた歴史と AI ブームの流れはほぼ一致している。一方で要約を AI で実施するという取り組み自体は古いものの，飛躍的な精度向上により，最近注目されている分野となる。

それらの AI に対する考え方を理解する上で，以下の「図 1 AI の歴史」に掲載している通り，1950 年代から現在にいたるまで，第一次 AI ブームから第三次 AI ブームがあり，それぞれ当時の環境が異なる背景もあることから，アーキテクチャも変遷をたどってきた。現在は Web サイトに存在するデータを中心とした大量の学習データと，それらを実際に学習するために必要となるマシンパワーの飛躍的な向上の両輪がうまく回る事で，生成 AI を中心とする第四次 AI ブームが到来していると言われたりもする。

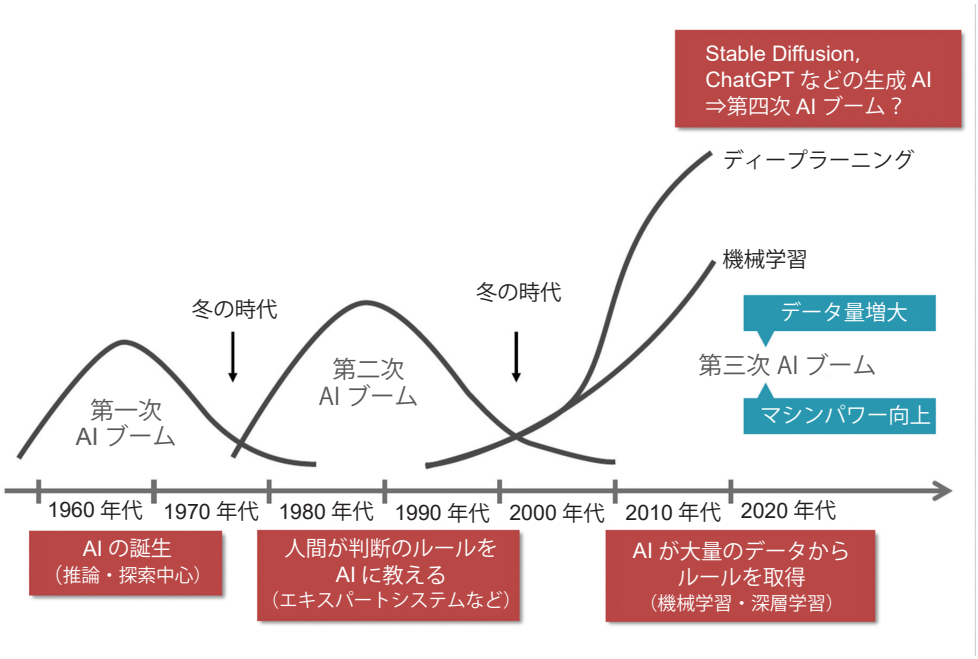

図 1　AI の歴史

3.　機械翻訳や AI 翻訳の精度向上

人手翻訳ではなく，何らかの機械装置を利用して翻訳を行うという概念は 1930 年から登場したが，その後様々な技術革新を経て，当初呼ばれていた機械翻訳から，人工知能が翻訳を行う，AI 翻訳へと進化を遂げた。

現在は，以下に掲載した「図 2 AI による自然言語処理と機械翻訳」にまとめた通り，第一次 AI ブームから第三次 AI ブームといった変遷の結果，機械翻訳の延長線上，あるいは発展した仕組みとして，機械翻訳と AI 翻訳も語られることが一般的である。

	自然言語処理	機械翻訳
第一次 AI ブーム：	文法などのロジックが基本	ルールベース機械翻訳
第二次 AI ブーム：	教師データ（コーパス）の機械学習	統計的機械翻訳
第三次 AI ブーム：	ディープラーニングの活用	ニューラル機械翻訳

図 2　AI による自然言語処理と機械翻訳

第一次 AI ブームでは，特定の構文しか翻訳する事が出来ない，ルールベースによる翻訳を行うことができる機械翻訳エンジンが登場する。しかしながら，実際の文章には様々な構文があり，ルールベースで全て包含されるものではないため，ルールから少しでも外れた構文の原文が来ると全く使い物にならないものであった。

第二次 AI ブームは辞書機能を有し，かつ文法の知識を機械に教えた，ルールベースの機械翻訳エンジンが登場する。この第二次 AI ブームでは，何らかの技能を AI に覚え込ませて AI に人の肩代わりをさせるという，エキスパートシステムが主流であったが，第二次 AI ブームで登場する機械翻訳エンジンも同様のアーキテクチャの延長線上にある。つまり，人手による辞書データや文法の知識などを教師データとして学習させてゆくのだが，学習させるデータは原文と訳文がきれいに整理された翻訳データ（翻訳コーパス）が必要であるため，利用頻度が低い原文には対応が難しく，かつ辞書に無い単語や熟語が来た場合はお手上げ状態であった。しかしながら，原文によってはある程度の精度で訳文を作る事が出来たため，多少翻訳の足しになる程度のツールとしては利用された。

そのような状況において，第三次 AI ブームで一気に性能が向上した機械翻訳エンジンが登場する。その背景として，GPU を活用することによるハードウェアスペックの飛躍的な向上と，人手頼りだった教師データの準備から，物量に物を言わせた学習データをとにかく学習させれば精度が向上する仕組みとなる，深層学習が現実的なものとして利用されたことにある。したがって第三次 AI ブームで登場した翻訳エンジンのアーキテクチャは，NMT（Neural Machine Translation: ニューラル機械翻訳）と呼ばれ，機械翻訳ではなく AI 翻訳を行うエンジンと言われるようになった。

さらに，2022 年から始まったとされる，生成 AI が中心となる第四次 AI ブームでは，機械翻訳を行うために大量の学習データを学習させた専用モデルではなく，自然言語処理を行うために作られた LLM（Large Language Model: 大規模言語モデル）の中の 1 機能として，翻訳を行うことが可能となる AI が登場している。一方で生成 AI には，実際とは異なる回答をあたかも正解のようにコンテンツを生成する，ハルシネーション（AI による幻覚）という大きな問題があるため，LLM を活用した AI 翻訳の訳文に対しても注意を払う必要があるが，それは従来から完璧な訳文を生成できずに，ある程度の精度の訳文を生成してきた機械翻訳～ AI 翻訳の歴史から見ると，さほど違和感は無いような気はするが，いかがだろうか。

4. NMT を活用した翻訳

翻訳作業を実施する際は，担当者や翻訳業者が対応できる言語セットや，保有する翻訳スキルの問題で，原文の言語と訳文の言語との関係が決まってくる。一方，NMT を活用した AI 翻訳の場合は，対象の言語セットに対する学習データの分量に比例した形で，AI の翻訳精度が異なるため，日本国内から見た場合，技術文書をはじめとする多言語展開は，精度面を考慮した場合は一般的に以下のように進められることが多くなる。

- 2 バイトコンテンツ（中国語，韓国語など）⇒ 日本語コンテンツを原文とした形で訳文を作成
- 1 バイトコンテンツ（フランス語，イタリア語など）⇒ 英語を原文とした形で訳文を作成

つまり，求められる訳文やターゲットとする言語を考慮した上で，原文を選択した方がより高い精度の AI 翻訳結果を得ることが出来る。特に，マニュアルや取扱説明書などの場合，多言語展開を行う場合がしばしば見られるので，そのような場合は基本的に日本語と英語のマニュアルをセットにした形で原文として用意し，それぞれの言語対応を進めていくといった手法が取られることが理想的である。

5. 翻訳支援ツール（CAT）

NMT の性能向上が著しいとはいえ完璧ではなく，原文の特性や記述方法などにより誤訳は発生する。それらをカバーするツールとして，CAT と呼ばれるツールを活用することで，翻訳業務を効率化させることができる。CAT はデスクトップツールやクラウドで提供されるものをはじめ，商用の製品が中心となるが，オープンソースのものも存在する。一般的な CAT は複数の NMT を選択することが可能であり，原文と訳文のセットごとに精度が高い NMT を個別に選択することが可能である。

6. PE（Post Edit）

人手翻訳でも完璧ではないのと同様，機械翻訳はもとより，AI 翻訳も万能ではなく，ある程度の誤訳が生じる事は致し方ない。翻訳者が誤訳部分を確認し補正する業務は PE（Post Edit）と言われるが，PE が出来る翻訳担当者は原文と訳文それぞれの言語に精通している必要があり，PE 作業を行う担当者には一定レベル以上の翻訳スキルが求められる。原文を機械翻訳や AI 翻訳にかけて訳文を生成するまでの時間は人手翻訳と比べると圧倒的に短時間で済むため，それだけ訳文は大量に発生することになる。したがって，機械翻訳や AI 翻訳が使われるシーンが増えれば増えるほど，PE 作業は右肩上がりで増加することになる。現時点においても，翻訳業界や翻訳担当者にとって，平易な翻訳は機械翻訳や AI 翻訳に取って変わられるために必要性が無くなる可能性が見込まれている。また，専門性の高いコンテンツや，誤訳が多い機械翻訳や AI 翻訳の言語セットの場合は PE 作業の必要性が高いため，技術文書においても PE 作業の必要性が高くなる。また，それ以外のコンテンツに関しても PE 作業に対するニーズは増加する傾向にあり，PE 作業に対応できる翻訳会社や翻訳者が生き残るのではないだろうか。

7. 翻訳メモリ（TM:Translation Memory）

特にマニュアルに見受けられるが，NMT や CAT，PE などによりせっかく完成した訳文が，製品のバージョンアップなどにより，原文を改訂しなければならないことが頻繁に発生する。そのような場合，改訂したマニュアルを NMT などで再度全翻訳してしまうと，改訂を行っていない箇所も再翻訳してしまうことになり，旧版で PE 作業などを通してせっかく仕上げた訳文の精度が落ちてしまう。したがってマニュアルなどの場合，改訂箇所のみを翻訳するといった手法である差分翻訳という方法が取られることがある。さらに改訂箇所の中で過去に翻訳した原文と訳文のパターンが保存されており，原文と訳文の内容が一致する場合は，その翻訳セットをそのまま活用することで翻訳自体の手間を省くことが可能となるが，原文と訳文のセットを保存して再活用するための機能を翻訳メモリ（Translation Memory）と言い，標準的な CAT ツールにはいずれも存在する機能となる。翻訳メモリは，技術文書の翻訳が繰り返されるほど，翻訳パターンはデータとして蓄積されるので，原文と訳文のセットに一致する可能性も高まることから，翻訳作業をより効率化することが可能となる。ちなみに翻訳メモリのデータ自体は，たいていの CAT ツールからは取り出せるとともに，相互の CAT ツールでも互換性がある場合がほとんどである。

8. 技術文書の作成時に注意すべきこと

機械翻訳や AI 翻訳の活用を念頭に置いた技術文書を作成する方法として，トピックライティングという手法があり，文章を部品化することで前後のコンテンツや文脈（コンテクスト）の影響を極力排除し，コンテンツの再利用性も高めることができる。したがって，前後の関係性が重要となる小説などのコンテンツには適さないが，技術文書の場合は書くべき内容が章節項の単位で構造化されており，目次構成に即した形でコンテンツを作成してゆくことが出来るため，トピックライティングは技術文書のライティングに適した手法として使われることも多い。

トピックライティングで書かれた文章は，それぞれの部品化されたコンテンツで閉じた意味を持つことになるため，トピック単位で翻訳をすることが可能となる。したがって，機械翻訳や AI 翻訳とは非常に相性が良く，高い精度で訳文を生成することができる。トピック単位で部品化されたコンテンツは，NMT による機械翻訳や AI 翻訳を手軽に行う事が出来るため，翻訳精度の確認とともに，原文が独立したトピック単独でもしっかり書かれているかを確認するために，あえて訳文から原文を再翻訳するという，リバース翻訳という手法で原文コンテンツの内容確認を行うという手法もある。

9. 技術文章を効率的に管理する CMS や CCMS

トピックライティングによりマニュアル等の技術文章を作成した場合の物理データ数は，ページ数・言語数・バリエーション・バージョンに比例する形で増加する。それらの物理データをファイルベースで管理することも可能だが，技術文書の運用を進めるにつれて物理データは飛躍的に増加するため，あまり現実的ではない。たいていの技術文書は製品のライフサイクルに応じた形，あるいは法改正などが行われた際に改訂が必要となるが，多言語で展開している技術文書の場合，改訂内容は全ての訳文に反映させる必要があるため，物理データをファイルベースで管理している場合，改訂時のヌケモレや先祖返りといったことが頻発する。

技術文書にとって致命的となるような，それらを防ぐ手法としては，技術文書をデータベース化するという手法が有効的である。データベース化した技術文書を管理するシステムは，CMS（Contents Management System）もしくは CCMS（Component Contents Management System）と呼ばれ，海外製品が多いが，日本企業向けに適した国内製品もいくつか存在する。CMS や CCMS は高額な海外製品が多く，また海外製品は国内向けサポートが余り期待できないといった問題も存在する。一方で国内製品は海外製品に比べてリーズナブルなものが多く，国内企業向けのサポートが手厚いベンダーも多い。

技術文書を扱う CMS や CCMS は Web CMS とは異なり，サーバによる外部への情報公開を行う目的ではないため，改訂管理や多言語コンテンツ管理に限らず，PDF や HTML へのワンソース出力や，部品化によるコンテンツの再利用など，様々な機能を有しているものが多い。それらの様々な機能に対応するため，データフォーマットは柔軟性が高く，ドキュメントのデファクトスタンダードである XML（Extensible Markup Language）ベースで管理されることが大半である。

10. 技術文書の要約

技術文書に限らず，要約作業は生成 AI が効果を発揮する分野となる。従来は入力文書から文字列を抜き出すという考え方がベースとなる「抽出型要約」が基本であったが，第三次 AI ブームでは，入力文書をエンコードした後で要約文書を作成する「抽象型要約」という手法で飛躍的に性能が向上した。単にプロンプトからコンテンツを生成するのではなく，入力するための要約元コンテンツが存在するので，ハルシネーションも発生しにくいとされているが，入力文書の内容や言語などによっては，元の文書に無い内容を生成したりする事もあるため，技術文書に求められる正確性を担保するという要件から考えると，用途にもよるが人によるファクトチェックは必須といえる。

逆にファクトチェックの手間が軽くても許容される技術文書の活用例としては，技術要素は必要だが製品の誤操作を引き起こすといった危険性が無いコンテンツ，たとえば SEO 対策などの Web マーケティングで使われるような，ブログ系 Web サイトで公開される技術コラム，あるいは論文などの技術文書から要約文を生成した上で手早くリサーチをするといった作業において，生成 AI の活用は非常に向いている。

11. 生成 AI を活用する上での注意点

生成 AI を活用する場合，対象となる学習データや自然言語処理技術の特性により，向き・不向きがあることを理解しながら活用すべきである。

特に，ChatGPT に代表される英語圏のベンダーが提供する大規模言語モデルは，英語版ウィキペディアなどをはじめとする様々な英語圏の Web コンテンツを主に学習しているという特性から，それらの大規模言語モデルは次のような性格を有している。

・英語圏で頻繁に登場する一般情報には強い

例）国の首都や地名，有名な政治家やハリウッド俳優の名前に関連した情報など

・比較的新しい情報には強い（ゆえに IT 関連の情報は精度が高い）
　例）サーバ，クラウドコンピューティング，GPU といった語句の説明など

また，以下のような英語圏以外に関する情報はハルシネーションが起きやすいとされている。
・各地域に根差した，ローカルの情報に依存する内容
　例）地元で有名な飲食店や施設，欧米ではあまり知られていない歴史に関する内容など
・国や地域に根差した独特の文化やルール，趣味や嗜好に影響される内容
　例）本将棋に関する内容，好きなラーメン店など

ただし，大規模言語モデルはすさまじいスピードで進化しているため，いずれの場合も近い将来は改善されていると思われる。

12.　翻訳や要約に生成 AI を活用する注意点

現時点で ChatGPT などの一般的な生成 AI を利用する際は，利用規約上は重要情報の外部流出が無い状態（オプトアウト）を選択することが可能であり，入力情報が大規模言語モデルに学習・記録されることを防ぐことができるため，個人情報や機密情報などが漏洩するリスクを軽減することができる。ただし規約上の話ではあるので，実際のところどうなっているかは生成 AI サービスを提供するベンダー側の処理次第となるため，秘匿性の高いデータを入力データとして提供しても良いのか，得られるメリットとの兼ね合いで慎重に判断する必要がある。

13.　まとめ

技術文書は正確性が求められるため，例えば技術文書のライティングにハルシネーションが頻発する生成 AI を活用することは，現時点においてまだまだ慎重に考えるべきではないだろうか。しかしながら，AI 分野の技術革新はすさまじく，現在の問題点はいずれ凌駕される時代が来ることは容易に想像できる。それゆえ，自律的な人工知能が自己フィードバックを繰り返すことによって，人間を上回る知能を得るのではないかと言われているシンギュラリティ（技術的特異点）は 2045 年に来るとされているが，ここ数年の人工知能の進化を目の当たりにすると，かなり前倒しになるのではと想像する。