医薬分野個人翻訳者による、医薬分野AI翻訳サイトの開発

合同会社クリニカルランゲージ　代表　山本隆之

【背景】

・現在医薬翻訳では、特に製薬会社向けの翻訳においてMT＋ポストエディットの使用が広がっている。

・発表者も現在は通常の人手による翻訳とポストエディットの仕事を行っている。多くの個人翻訳者は、今後AIによって人間による翻訳の仕事が無くなることを懸念しており、発表者も同様に考えていた。

・一方で、ポストエディターとして仕事をするうちに、最近のディープラーニングを基盤技術とするニューラルMTの進化を非常に感じるようになり、AIに興味を持つ。

・新しいことにチャレンジするという意味で、AI専門のプログラミングスクールを受講したところ、ある程度翻訳サイトのウェブアプリを自作できることが分かった。

【目的】

自分の翻訳専門分野である医薬翻訳分野でAI翻訳サイトを開発する。完成できるかどうか分からないが、自作でどこまでできるかやれるところまでやってみる。

【方法】

①AIプログラミングスクールの課題であった、花の種類の判別アプリ（花の蕾と花弁の大きさから花の種類を判定する）と物体認識アプリ開発の本を基に、Google AutoML TranslationのAPIコードを組み込む。

（佐藤 昌基 、平田 哲. Python FlaskによるWebアプリ開発入門 物体検知アプリ&機械学習APIの作り方. ）

②Google AutoML TranslationはGoogle翻訳をベースに対訳で追加学習することで、ドメイン特化のモデルが作れる。そのため、公表されている対訳を集める。

翻訳サイト作成にに必要なこと

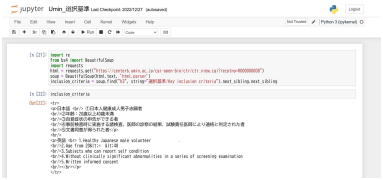
・サイト自体の開発（本と他社サイトから、見よう見まねでやってみる）

・対訳データの収集

・用語集の作成

インターネット上のデータをpythonのコードでスクレイピングまたは、PDFをエクセル上の対訳にする。

（TMXファイルの作成）



インターネット上のデータのスクレイピング

使用した技術スタック（主なもの）

・Flask	・Google Translate API	
・Stripe（決済システム）	・Google Cloud Run	
・Cloud SQL	・ChatGPT API	
・PubMed API	・Clinical.Trials.gov API	など

途中過程

・医学や治験翻訳の対訳は、公表されている医薬品開発ガイドラインなどを集めGoogle AutoMLでトレーニングを行った。

・モデルをトレーニングするために対訳を集めることに時間がかかり、通常のGoogle翻訳APIを使ってアプリの体裁を作っていた。

開発中の画面

ChatGPTの登場

2022年11月にChatGPTが登場したことにより、プログラミングの速度が上がった。特にデバッグが有用で、エラーコードの特定に役立った。

用語集作成など簡単なアプリであれば、ChatGPTが瞬時に作ってくれた。（これを自分のアプリに組み込む作業を行った）

ChatGPTに相談しながら開発

【結果】

初公開：2023年10月

この時はGoogle AutoML Translationベース。翻訳品質はいまいち。

通常のGoogle翻訳はすでに解剖用語などを正しく出力し、流暢に翻訳できるが、対訳でトレーニングすることでかえって文が途切れたりしていた。対訳を増やすとBleuスコアが低下した。

翻訳品質には納得できなかったがとりあえずオープン。

2023年10月サイトオープン

原文	訳文	Bleu	ベース Bleu	Bleu の増加	ベースモデル	対訳数
日本語	英語	37.748	22.134	15.614	Google NMT	14,506
英語	日本語	45.817	25.697	20.119	Google NMT	14,736
日本語	英語	29.797	22.163	7.634	Google NMT	111,654

Bleuスコア

この1年のChatGPTの進化

Google AutoMLの翻訳では、文が途切れることが多かったため、ChatGPTによる校正ボタンを付けていた。しかし、当時GPT4の価格は高かったため、GPT3.5にしていたが、GPT3.5はである調「常体」で和訳することができなかった。

その後、2024年7月にGPT-4o miniがリリースされ、常体・敬体の修正が正しく反映できるようになった。

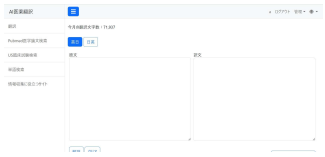
現在の画面と機能

1. AI翻訳
2. プレインランゲージで翻訳
3. PubMedのAPIから医学論文を検索・翻訳
4. ClinicalTrials.govのAPIから臨床試験を検索・翻訳
5. 医学分野用語集
6. 英語ページも作成

ほぼベースはChatGPTのプログラムで、つなぐことを自分がやっていた。

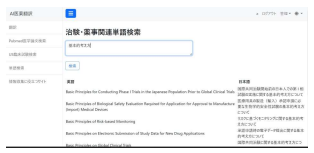
現在は、通常のGoogle翻訳に用語集を使用し、ChatGPTで修正する形で翻訳している。

現在の画面



PubMed翻訳画面

医学・薬事用語集画面



【考察・結論】

・プログラミングの経験ゼロから始めて3年かかったが、最初にイメージしていたものがほぼできた。

・2022年11月にChatGPTが登場してから、翻訳もプログラミングも非常に精度と速度が上がった。

・1人でできないことは、大量データの前処理など、単純だが労力の必要なデータ作成に関わることであった。

・今回自作したサイトに近いものとして、PubMed論文を動画にまとめるサイトがあり、元MRの方が開発されている。

・翻訳もテキストからテキストだけではなく、テキスト→音声翻訳、テキスト→翻訳動画など、モダリティが変わってくるものが出てくる。

・生成AIで翻訳・プログラミングの難易度が下がったので、翻訳者・翻訳会社（機械翻訳会社も含め）以外でも翻訳エンジンを作ることができる。

・機械翻訳エンジンは、今後ChatGPTのfine-tuningやプロンプトを組み合わせれば誰でも作成できる。DeepLも医療分野の専用エンジンを出しており、GAFAM以外の通常の企業では、エンジンに差は付けられないのではないか？