

ニューラル翻訳を用いた中国特許機械翻訳精度の検証:

中国特許の日本語及び英語への機械翻訳精度の検証

○田畑文也¹⁾

富士フイルム(株)¹⁾

〒421-0396 静岡県榛原郡吉田町川尻 4000

E-mail: fumiya.tabata@fujifilm.com

Study of Neural Machine Translation accuracy of Chinese Patents:

Study of Machine Translation accuracy of Chinese Patents to Japanese and English

TABATA Fumiya¹⁾

FUJIFILM Corporation¹⁾

4000, Kawashiri, Yoshida-cho, Haibara-gun, Shizuoka , 421-0396 Japan

E-mail: fumiya.tabata@fujifilm.com

【発表概要】

AI(人工知能)の急速な進歩に伴い、自然言語処理技術が急速に進化を遂げている。これにより、機械翻訳のアルゴリズムについても、ニューラル翻訳(NMT)が登場した。また特許調査については、中国特許の急激な増大により、中国語で書かれた特許を査読する頻度も多くなり、その際には、まず日本語または英語に機械翻訳したもので、内容を理解しようとするのが通常である。しかし、現状では中国特許の機械翻訳の精度は高くないことも多い。中国特許についてNMTを用いて翻訳した場合、その翻訳精度を調べた。ただし、本稿を記した時点(2017年9月)では、中国語から日本語へのNMTに対応したものは少なく、中国語から英語へのNMTシステムも合わせて調べることにより、翻訳精度を評価した。

評価した結果、従来のルールベース翻訳や、統計翻訳とNMTを比べると、NMTの方が、単語レベルでの翻訳精度では必ずしも高いとは限らず、むしろ翻訳精度が低下する場合もあることが分かった。しかし、文としての意味を理解する上では、NMTの方が文の構成を理解し易い傾向があり、これについては発表までに詳しくまとめる予定である。また、中国語から日本語への翻訳より、中国語から英語への翻訳のレベルの方が高く、翻訳精度を求めるならば、日本語より、英語で見る方の効果の方が現時点では大きいことが分かった。

【キーワード】

機械翻訳, ニューラルネットワーク, ニューラル翻訳, NMT, 中国特許

1. はじめに

近年、AI(人工知能)の急速な進歩に伴い、自然言語処理技術が急速に進化を遂げている。これにより、機械翻訳のアルゴリズムについても、設定したルールに基づいて翻訳するルールベース翻訳(RBMT)から、統計的な解析を用いた統計翻訳(SMT)、そして機械学習を用いたディープニューラルネットを使用したニューラル翻訳(NMT)とトレンドとしては移行している。(表 1)

表 1. 機械翻訳のアルゴリズム

翻訳方法名	翻訳のアルゴリズム
ルールベース翻訳(RBMT)	ルールに基づいて翻訳する方法
統計翻訳(SMT)	大量の対訳データを解析し、その統計結果から適した訳し方を割り出す翻訳方法
ニューラル翻訳(NMT)	機械学習を用いたディープニューラルネットを使用し、単語の意味だけでなく接頭辞や語幹、単語の位置なども考慮し、自然な文の流れを分析して翻訳する方法

ここで、知財実務では、原文が英文で書かれた特許よりも、莫大な件数が出願されている中国語特許の調査および査読に膨大な労力がかかっている場合があるのも現状である。しかし、現状の中国特許の機械翻訳の精度は高くなく、この問題に対処する必要がある。

機械翻訳については、2016 年頃より、Google 翻訳が、英語から日本語への翻訳アルゴリズムを NMT に切り替えるなど、徐々に NMT に対応は進み始めているが、本稿執筆時点(2017 年 9 月)では、まだ NMT に対応したものは少なく、かつ中国語から日本語へ対応したものはさらに少ない。このように、現在は過渡期であるが、中国特許の日本語および、英

語への機械翻訳の精度を検証し、どのような機械翻訳システムの翻訳精度が高いかを調べたので報告する。ただし、予稿集の時点ではデータとしては、単語レベルの評価とし、本発表にて、さらに文としての評価についても述べる予定である。

2. 評価方法

2.1 評価に用いた翻訳システム

現時点では中国語から日本語への機械翻訳システムで、NMT に対応しているのは、筆者の知る限り、Google 翻訳、MS(マイクロソフト)翻訳のみで、これにレファレンスとして、日本特許庁が提供する JPO 中韓文献翻訳・検索システム(以下、JPO 中韓文献)をあわせて評価した。(表 2、図 1~図 3)

なお、Google 翻訳は 2017 年 8 月下旬に、中国語から日本語への翻訳アルゴリズムが、SMT から NMT(推定)に切り替えられたが、評価したデータのうち、SMT のデータも一部残っており、SMT の結果も併記した。

表 2. 今回評価した中国語から日本語への機械翻訳システム

システム	URL	中国語→日本語翻訳アルゴリズム
Google 翻訳	https://translate.google.co.jp/	NMT(推定)
MS 翻訳	https://translator.microsoft.com/neural	NMT および SMT
JPO 中韓文献	http://www.ckgs.jpo.go.jp/	RBMT



図 1. Google 翻訳

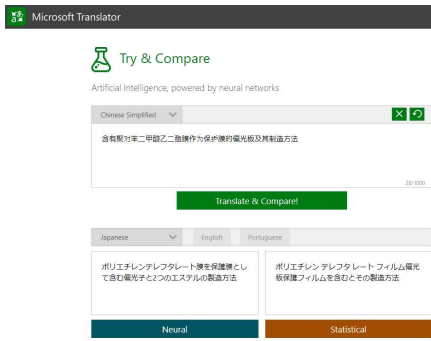


図 2. MS 翻訳

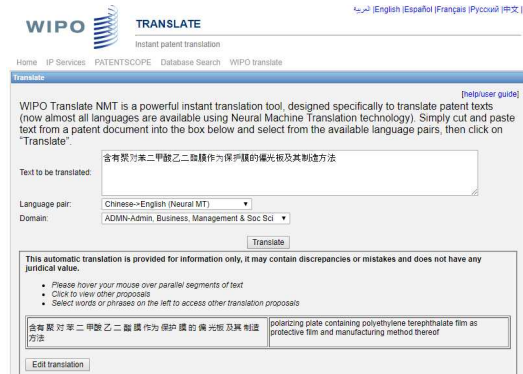


図 4 WIPO 翻訳

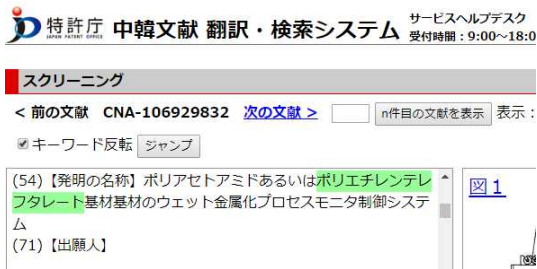


図 3. JPO 中韓文献翻訳・検索システム



図 5 百度翻訳

また、中国語から英語への機械翻訳については、Google 翻訳、MS 翻訳の他に、WIPO (世界知的所有権機関) が提供する WIPO 翻訳、および中国の百度(Baidu)社が提供する百度翻訳についても合わせて評価した。(表 3 および図 4、図 5)

表 3. 今回評価した中国語から英語への機械翻訳システム

システム	URL	中国語→英語 翻訳アルゴリズム
Google翻訳	https://translate.google.co.jp/	NMT
MS翻訳	https://translator.microsoft.com/neural	NMTおよびSMT
WIPO翻訳	https://patentscope.wipo.int/translate/translate.jsf?interfaceLanguage=en	NMTおよび従来型(非公表)
百度翻訳	https://fanyi.baidu.com/	NMT(推定)

2.2 評価に用いた技術用語

表 4 に示す計 3 種の技術用語について、中国特許・実案の発明の名称で各 10 件ずつ調べ評価した。

表 4. 今回評価した技術用語

日本語	中国語	意図
ポリエチレンテレフタレート (PET)	聚对苯二甲酸乙二酯	確立された汎用用語
シェールガス	页岩气	比較的新しい用語
クラッシュボックス (衝撃吸収ボックス)	碰撞吸能盒	業界専門用語的で、あまり一般用語ではないもの

2.3 翻訳精度評価基準

日本特許庁の特許文献機械翻訳の品質評価手順¹⁾を参考に、技術用語を表 5 に示す基準で、原文の中国語を理解できるもの (N=1 人) が、独自に点数評価し、各 N=10 の結果を平均化したもので評価した。

表 5. 技術用語の翻訳精度基準

ランク	レベル	点数
A(適訳語)	人手翻訳に照らし、技術的に同義かつ一般的に用いられる訳語である。	2
B(可訳語)	技術用語として一般的に用いられる訳語ではないが、意味はおおむね正しい。	1
C(誤訳語)	誤訳である。	0
D(不訳語)	未知語、訳漏れである。	0

3. 結果

3.1 中国語から日本語への機械翻訳

3種(各 N=10 平均)の技術用語について、中国語から日本語への単語としての機械翻訳の精度を評価した結果を図 6 に示す。

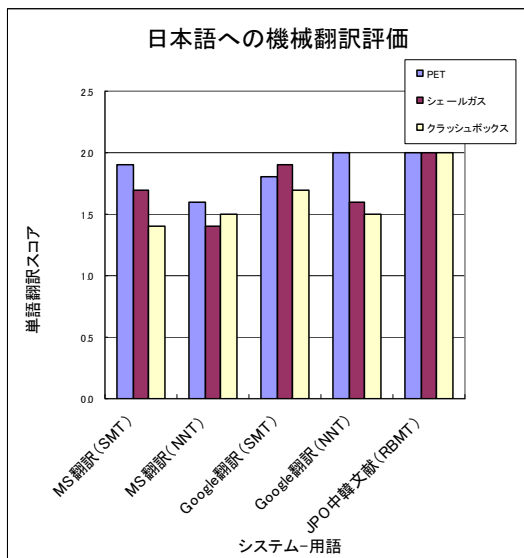


図 6 中国語から日本語への翻訳精度評価

翻訳アルゴリズムが RBMT の JPO 中韓文献が、最も翻訳精度良く、従来の翻訳アルゴリズムのものより、むしろ NMT 方が翻訳精度悪い結果である。

3.2 中国語から英語への機械翻訳

同様に中国語から英語への翻訳精度を評価した結果を図 7 に示す。

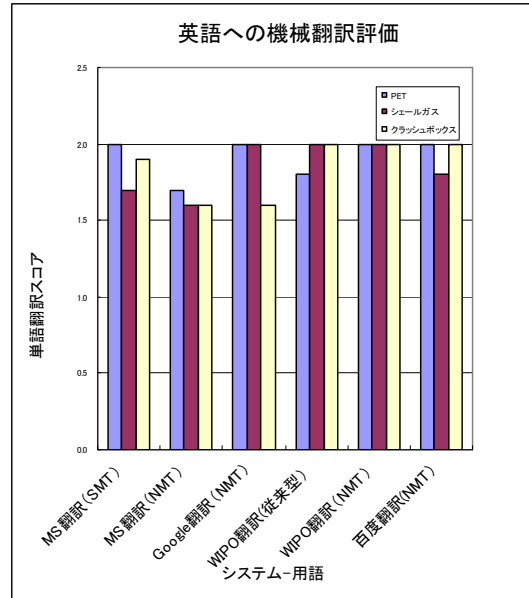


図 7 中国語から英語への翻訳精度評価

中国語から英語への機械翻訳は、全体的に日本語への機械翻訳よりレベルが高く、翻訳アルゴリズムの差より、システムの差の方が顕著である。

4. 考察

中国語から日本語への翻訳については、翻訳アルゴリズムが RBMT の JPO 中韓文献が、最も翻訳精度良く、同じシステム同士の SMT-NMT アルゴリズム比較でも、NMT が必ずしも高い訳ではなく、むしろ今回の評価では平均値として、SMT に劣る。これは、翻訳アルゴリズムの影響で、NMT より、RBMT や、SMT の方が、技術用語の訳を定義した場合、それが反映されやすいためと考える。また、技術用語による差もあり、やはり一般的に確立されている PET などの用語の方が、翻訳精度が高い傾向にある。

英語についても、同じシステム同士の比較では、NMT が必ずしも高い結果を示すわけではないが、全体として、日本

語よりも英語の翻訳精度が高く、翻訳アルゴリズムで比較するより、日本語ではなく英語に翻訳させる方が SMT、NMT とも翻訳精度向上への寄与は高いことが分かった。

ただし、単語の翻訳精度としての評価は本報告の通りであるが、評価の過程で文として見る場合、NMTの方が、文の意味が分かり易い傾向にあり、これについては、本発表までに評価を進め、発表する予定である。

また、現時点では、NMTは過渡期であり、以前評価した結果と、現時点で評価した結果が異なる場合もあり、急速に進化している模様である。

5. おわりに

最後に、本報告は2017年度の「アジア特許情報研究会」のワーキングの一環として報告するものであり、会のメンバーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

6. 参考文献

[1] 日本特許庁 情報技術統括室,
特許文献機械翻訳の品質評価手順
Ver1.0 (平成 26 年 6 月),
http://www.jpo.go.jp/shiryuu/toushin/chousa/pdf/tokkyohonyaku_hyouka/01.pdf (accessed 2017-09-12)