特許翻訳システムの社会実装を目指したJBMIAの取り組み

JBMIA's efforts toward social implementation of patent translation system

一般社団法人ビジネス機械・情報システム協会　知的財産委員会

金子　清隆

2010年より一般社団法人ビジネス機械・情報システム協会知的財産委員会の委員。2014年より同委員会幹事会委員、2021年より同委員会幹事会オブザーバー兼機械翻訳WGアドバイザー。現職は、コニカミノルタ株式会社知的財産部部長付シニアアドバイザー。

kiyotaka.kaneko@konicaminolta.com

1　はじめに

人工知能（AI）を用いたAI翻訳[1]は、従来の機械翻訳と比べて流暢で精度の高い翻訳が可能になった。

特許情報プラットフォーム（J-PlatPat）[2]や、世界特許情報全文検索サービス[3]などに実装され、外国公報の調査効率化・容易化に貢献している。

一方、高い専門力を有する翻訳者が担っている外国出願時の明細書翻訳は、AI翻訳でも精度不足と言われ、実用化に対する壁が引き続き存在している。

JBMIA（一般社団法人ビジネス機械・情報システム産業協会）知的財産委員会は、日本企業や研究機関にとって共通の課題である特許明細書の翻訳コストおよび翻訳時間を大幅に削減する特許翻訳システムの実現を目指し、AI翻訳を開発されている外部機関と協働してAI翻訳の性能向上への取り組みを進めている。

本稿では、JBMIA知的財産委員会による取り組みを紹介するとともに、AI翻訳エンジンをドメイン適応によって高精度化したポイントについて述べる。読者の方々の参考になれば幸いである。

2　JBMIAとは

はじめにJBMIAについて簡単に紹介する。

JBMIAは、1960年に設立された日本事務機械工業会をルーツに持ち、複合機、プリンター、プロジェクターなどのビジネス機器を製造・販売する企業各社が会員となって様々な活動を展開している産業協会である。知的財産委員会には、委員総数88名（2022年6月現在）が参加して活動を展開しており、特許翻訳システムを研究するための機械翻訳WG（10社21名）を含む（図1）[4]。

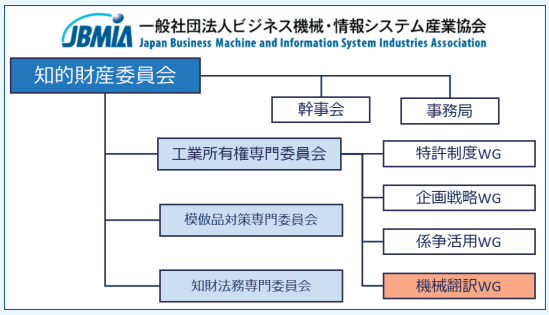

図1　JBMIA　知的財産委員会

3　特許翻訳システムの実現目標

JBMIAが考える特許翻訳システムでは、AI翻訳エンジンに対して、明晰な日本語原文を高精度に翻訳可能とすることを目標として設定した。翻訳者の読解力によってのみ翻訳が可能な日本語原文は、原文を改善することでAI翻訳エンジンを活用することとしている。（図2）

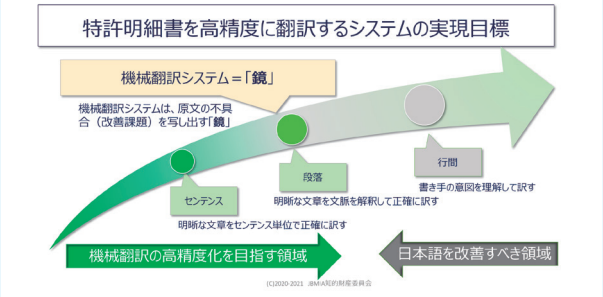

図2　特許翻訳システムが高精度化を目指す領域

高精度な AI 翻訳エンジンの実現には、良質な教師データが大量に必要となる。最初のステップでは、知的財産委員会の会員企業が保有する日英対訳コーパスを最低 100 万センテンス集めて、AI 翻訳エンジンをドメイン適応（アダプテーション）して構築することとした。

次のステップでは、用語統一し明晰な日本語（産業日本語）で記載された原文とその訳文 100 万センテンス以上を教師データに用いて、AI 翻訳エンジンをドメイン適応して構築することを目指す。

最終ステップでは、用語統一し産業日本語で記載された原文とその訳文だけで深層学習した AI 翻訳エンジンを構築することを目指すとした（図 3）。

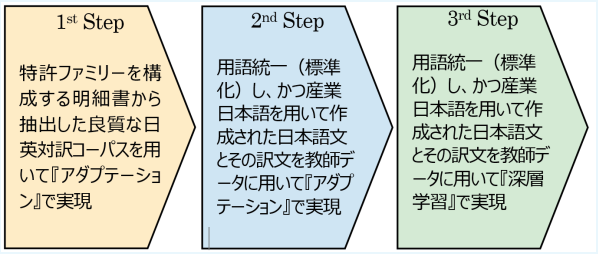

図 3　高精度 AI 翻訳エンジン実現のためのステップ

第 2 ステップで用語統一と産業日本語を採用する理由は以下の通りである。個社で良質な日英対訳コーパスと判断したものであっても、それらを集めた場合に必ずしも良質な日英対訳コーパスとはならない。具体的には、用語や訳語の選択に関して、業界としての共通ルールが存在していない。そのため、個社では用語や訳語に揺らぎがない日英対訳コーパスであっても、各社から集めた日英対訳コーパス全体では、用語や訳語の揺らぎが生じている（図 4）。このことは、AI 翻訳の訳文を評価する際にも、同じ訳文に対して各社で異なる評価結果を招いてしまう。そのため、高精度な AI 翻訳を実現するために用語統一は必須と考えている。また、産業日本語を採用する目的は、文章の理解容易性・明晰性の観点から主語や目的語を明示することなど、特許明細書をローコンテクストな文章とすることにある。産業日本語に関しての詳細は特許ライティングマニュアル [5] 等を参照されたい。

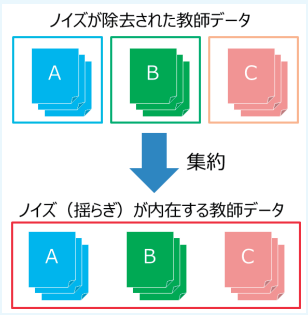

図 4　教師データ（対訳コーパス）を集約した時の課題

なお、AI 翻訳サービスも国際的な市場シェア争奪競争になると予想される。このため、特許翻訳システムは図 5 に示すように、官民の協働で日本の公共インフラとして社会実装することが好ましい。明瞭・明晰な日本語文による特許明細書の作成、良質・大量な教師データの提供、用語･訳語の標準化を企業が協業して取り組み、プラットフォームについては、公共財として開発・運用されることが望ましい。

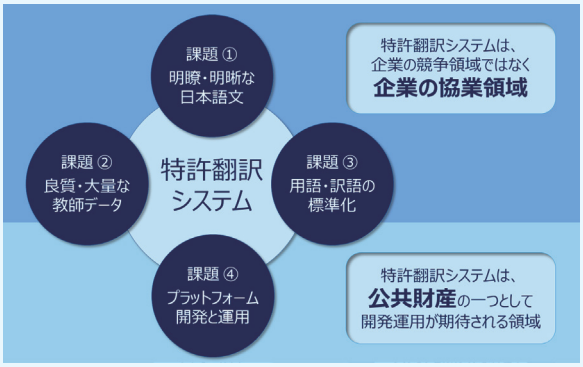

図 5　特許翻訳システムの社会実装

4　高精度 AI 翻訳エンジンの構築

これまでに構築した AI 翻訳エンジンについて説明する。なお、翻訳エンジンの性能評価は、英訳文を米国特許出願明細書に採用した場合の実務評価とし、表 1 に示した評価基準に基づいて知的財産委員会の機械翻訳 WG に参加している各社の知財部員が実施した。本稿中でも AI 翻訳エンジンの性能比較評価に本評価基準を用いている。

表 1　翻訳文の評価基準

判定	定義	米国出願後の対応
①	訳文をそのまま使用可能	影響なし
②	軽微な修正が必要※修正せずとも意味は通じる	必要に応じて補正
③	軽微でない修正が必要※意味が通じず修正必須	誤訳訂正

4.1　アダプテーションエンジン

2019 年 12 月から国立研究開発法人情報通信研究機構（NICT）と共に、図 3 の第 1 ステップに相当する高性能な AI 翻訳エンジンの構築に関するフィージビリティ・スタディを実施した[6, 7]。知的財産委員会が 100 万センテンスの対訳コーパスを提供し、NICT が特許 NT エンジンをドメイン適応して JBMIA 専用アダプテーションエンジンを構築した。図 6 に結果を示すが、高精度化（ドメイン適応）の効果は得られず、想定と異なる結果となった。

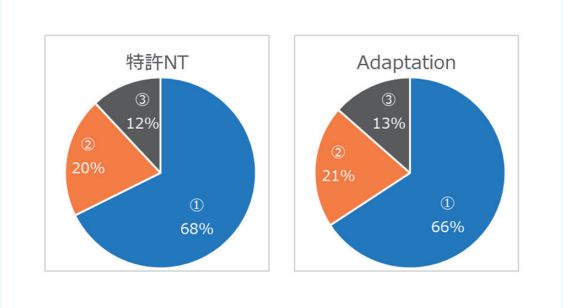

図 6　AI 翻訳エンジンの性能比較（1）

評価結果を詳細に検討すると、特許 NT エンジンでは誤訳となる原文でも、JBMIA 専用アダプテーションエンジンでは正しく訳出されるケースや、その逆のケースが観測された。即ち、2 つの AI 翻訳エンジンは、総合性能では大差がないが、得意 / 不得意な領域に相違があった。特許ファミリーを構成する明細書から抽出した良質な日英対訳コーパスであっても、『アダプテーション』に用いる教師データとしては品質が不足していたものと思われる。

4.2　マルチ NMT 翻訳

第 1 ステップの結果が想定と異なったことを受け、膨大な対訳コーパスの整備を必要とする第 2 ステップの実施を中止した。一方、NICT から提案されたマルチ NMT 翻訳について評価を行うべく、日本特許翻訳株式会社によって提供される JBMIA 専用ニューラル自動翻訳サービスを 2021 年 3 月より利用開始した。

マルチ NMT 翻訳とは、特性の異なる複数の AI 翻訳エンジンで同時に翻訳を行い、最も高精度な訳文を自動選択して出力するものである。JBMIA 専用ニューラル自動翻訳サービスは、NICT の特許 NT エンジン、JBMIA 専用アダプテーションエンジン、JBMIA 専用アダプテーション +EBMT（Example-based Machine Translation）エンジン[8] を実装し、これら 3 つの AI 翻訳エンジンを用いてマルチ NMT の性能テストを実施した。

日本特許翻訳株式会社のマルチ NMT 翻訳システムは、期待通りの翻訳性能を示した（図 7）。マルチ NMT 翻訳システムを更に性能向上させるためには、短文領域で高い翻訳精度を持つ AI 翻訳エンジンと組み合わせることが最も有効であり、次に最適翻訳文を選択するアルゴリズム改善であることが判明した。

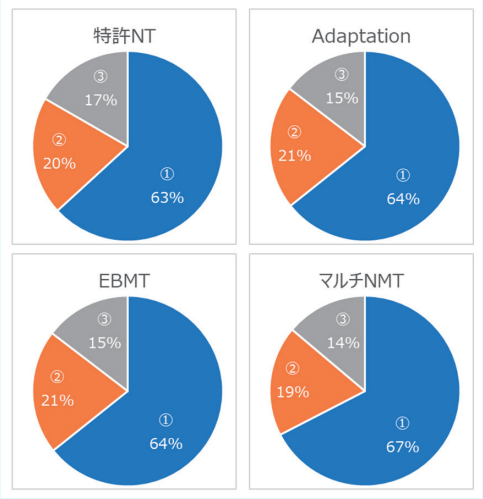

図 7　AI 翻訳エンジンの性能比較（2）

4.3　TM アダプテーションエンジン

少量の対訳コーパスでアダプテーションすることが可能となる、新しいドメイン適応手法が NICT から提供開始された。日本特許翻訳株式会社の JBMIA 専用ニューラル自動翻訳サービスでも、2021 年 5 月より「TM アダプテーション」というサービス名称で利用可能となった。

TM アダプテーションでは、数千センテンスの日英対訳コーパスでドメイン適応が可能になる。当初のアダプテーションでは、最低でも 100 万センテンス、望ましくは 1,000 万センテンスの日英対訳コーパスが必要とされており、各段に改善されている。

日本特許翻訳株式会社から TM アダプテーションエンジン構築の機会をいただき、同社の協力の下で機械翻訳 WG は 2021 年 10 月より TM アダプテーションを用いた AI 翻訳エンジンの構築と検証を開始した。

TM アダプテーションによって構築する AI 翻訳エンジンは、図 3 の第 2 ステップに相当する高品質な訓練

データを用いてドメイン適応を行うことで、高精度且つ予見性が高い翻訳文を出力可能となることの検証を目標に設定した。

予見性が高い翻訳文を出力可能とするためには、高精度な日英対訳コーパスが必要となる。そこで、後述する整備ガイドラインを設定し、機械翻訳 WG に参加する各社の電子写真分野の特許明細書から集めた 6,700 センテンスの日英対訳コーパス全てをガイドラインに沿って整備した。

このようにして構築した AI 翻訳エンジン（TM アダプテーションエンジン）の BLEU（Bilingual evaluation understudy）値は 88、RIBES（Rank-based Intuitive Bilingual Evaluation Score）は 96 と、夫々これまでにない高い数値となった（図 8）。

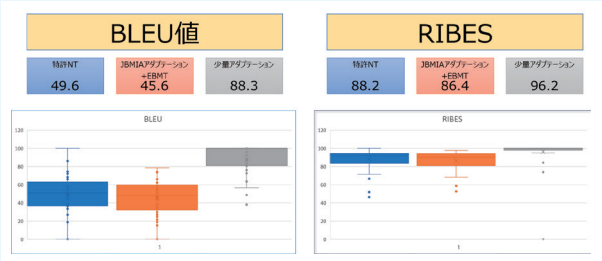

図 8　AI 翻訳エンジンの性能比較（3）

訓練データと同じドメインの原文 100 センテンスを用いて翻訳精度評価を実施したところ、特許 NT エンジンと比較して TM アダプテーションエンジンは、そのまま使用可能と評価される訳文（評価①）が 14 ポイント増えており、大幅な性能改善効果を確認した（図 9）。

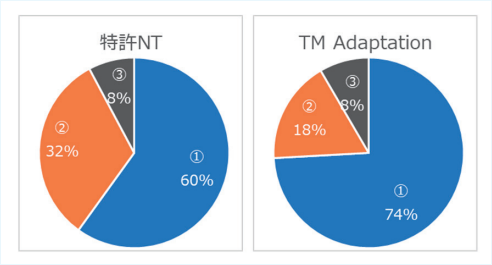

図 7　AI 翻訳エンジンの性能比較（2）

予見性の高い訳文を出力しているかを判断するため、原文に含まれる用語が常に同じ訳語で訳出される、所謂訳揺れのない訳文を出力可能であるかを評価した。用語の標準化を行った TM アダプテーションエンジンは、訳揺れが生じないことを実証できた。対訳コーパスに含まれる専門用語を統一するために作成した標準用語集では「画像形成部」の訳語を “image forming section” と規定している。テスト用原文として用意した 100 センテンス中、「画像形成部」を含む 13 センテンスの訳語を点検した（表 2）。特許 NT エンジンでは訳揺れが生じていた。TM アダプテーションエンジンでは 13 センテンス全てで、標準用語集で規定した “image forming section” で訳出され、訳揺れは生じていなかった。他の用語に関しても同様のテストを行ったが何れも訳揺れは確認されなかった。

表 2　訳揺れの抑止

「画像形成部」の訳語	特許 NT	TM Adaptation
image forming section	1	13
image forming unit	11	0
image forming portion	1	0

4.4　マルチ NMT の実用性

JBMIA 専用ニューラル自動翻訳サービスに、TM アダプテーションエンジンを追加したことで、新たな AI 翻訳エンジンを用いたマルチ NMT 翻訳を利用可能とした。まだ試行段階ではあるが、米国出願用明細書をマルチ NMT 翻訳とポストエディットによって作成したところ、全て人手翻訳した場合と比較して、半分の工数で完成させることができた。日本語明細書の原文品質や、TM アダプテーションエンジンが利用できるドメインであるか否かの影響は無視できないが、1 つの事例として紹介する。

4.5　アダプテーションエンジンと TM アダプテーションエンジン

TM アダプテーションエンジンは、少量の対訳コーパスにも関わらず、ドメイン適応効果を顕著に示した。この主因は、対訳コーパスの質の違いであると考えられる（表 3）。

表 3　ドメイン適応に用いた対訳コーパス比較

AI 翻訳エンジン	アダプテーションエンジン	TM アダプテーションエンジン
1. 対象ドメイン	個社の出願分野（個社任せ）	電子写真分野（指定）
2. 対訳コーパスの文章数	100 万	6,700
3. 訓練データの文章数	60 万	6,700
4. コーパス品質管理	なし（個社任せ）	実施
5. 訳語統制	なし（個社任せ）	あり（JBMIA 用語集）
6. 直訳と意訳	あり（個社任せ）	直訳のみ
7. 文字数制限	なし（個社任せ）	あり（原文 150 文字以内）
8. 英訳時の英文分割	あり（個社任せ）	あり（原文も分割修正要）
9. 明晰性	指定せず（個社任せ）	明晰でないものを除外

4.6　高品質な対訳コーパスの作成

高精度な AI 翻訳エンジンの構築方法として、TM アダプテーションは有効だと考えられる。そして、TM アダプテーションでは、対訳コーパスの整備が最も重要となる。

既に述べているが、一件の特許明細書の中では、原文も訳文も用語統一されている。しかし、各社から収集した対訳コーパスでは、各社毎に異なる用語で統一されているため、全体として原文と訳文ともに用語統一されていない。更に、米国特許明細書として好ましい英文とするために、意訳した訳文や原文で省略されている主語あるいは目的語を補った訳文も含まれている。これらは高精度な AI 翻訳エンジンを構築するための訓練データとして好ましいものではない。

そこで、訳揺れを抑え高精度な翻訳を実現するために、対訳コーパスのガイドラインを定めて全ての対訳コーパスを点検整備することが必要となる。TM アダプテーションエンジンを構築した際のガイドラインを表 4 に示す。

表 4　対訳コーパスのガイドライン

	ルール	対応
1	直訳となっていること	接続詞は接続詞ルールに従う。
2	原文と訳文の態は同じにすること	態が異なる場合は、原文と訳文の態を揃える。
3	原文および訳文において誤記がないこと	参照符号が不一致の場合は、原文に合わせて訳文を修正する。
4	小見出しは最初の用語のみヘッドキャピタルにすること（固有名詞で大文字を使用している場合は除く）	最初の用語以外に大文字が使用されている場合は小文字に変更する。
5	原文と訳文が各一文ずつになっていること	原文 1 文に対し訳文が 2 文以上になっている場合は、訳文に合わせて原文を分割する。
6	不要な括弧は含めない	訳す必要のない括弧は原文と訳文から削除する。
7	算用数字を使用しない（除外：参照符号、数値）	原文に含まれる算用数字は漢数字に変換する。
8	原文は全て全角とする	原文に半角が含まれている場合は全角に変更する。
9	カタカナ表記の末尾の長音記号（ー）は省略しない	長音記号（ー）が付いてない場合は、原文に追加する。
10	専門用語は指定された用語になっていること	専門用語のルールに従う。

4.7　TM アダプテーションの可能性

TM アダプテーションは、少量の対訳コーパスでドメイン適応（アダプテーション）を実施できる点が魅力となっている。一方で、訳揺れを抑止させるために必要な対訳コーパス数が明らかでなく、大規模な用語集を学習させるために大量な対訳コーパスが必要となる可能性がある。

また、TM アダプテーションエンジンで高精度に翻訳可能な対象領域（ドメイン）の拡張も課題の 1 つである。対象領域が狭いと数多くの TM アダプテーションエンジンを用意することとなり現実的でない。

機械翻訳 WG では、電子写真分野に続いてインクジェット分野で、TM アダプテーションの実施準備を進めている。既に整備を終えた電子写真分野の日英コーパスと、新たに整備するインクジェット分野の日英対訳コーパスとを用いて図 10 に示す 3 つの AI 翻訳エンジンを構成し、TM アダプテーションエンジンの性能を評価するとともに、ドメインの拡張性についての知見を得ることを予定している。

図 10　TM アダプテーションの拡張性評価

5　終わりに

知的財産委員会の機械翻訳 WG では、AI 翻訳の高精度化だけでなく、特許明細書の日本語原文を明晰化する活動（産業日本語の活用）も推進している。2021 年度までの成果を「日本語ルールブック」という形に纏めて、JBMIA 会員企業に公開した。今後ブラッシュアップを進めながら外部への公開も積極的に進めるとともに、日本語明細書作成に特化したオーサリングツールの実現に向けて関係部門との連携強化も進めたい。

JBMIA では、明晰な日本語文章と高精度な AI 翻訳を実装した特許翻訳システムによって、ポストエディットの負荷を大幅に軽減可能となることを立証し、他の産業分野の参考になることを目指している。そして、官民一体で産業日本語と高精度な AI 翻訳エンジンを備えた特許翻訳システムの普及を加速させ、日本企業の国際競争力強化を牽引するとともに、特許明細書の誤訳訂正等に対する柔軟な法的対応についても国際協調の下で実現が図られるよう働き掛けを行いたい。

最後になるが、本活動の推進にご協力をいただいている NICT、日本特許翻訳株式会社にはこの場を借りて改めて感謝を申し上げたい。また、日本企業全体のための活動として後押しをして下さっている第 6 代大崎委員長（元京セラドキュメントソリューションズ）、第 7 代真竹委員長（キヤノン）、第 8 代石島委員長（リコー）と、

機械翻訳 WG の活動を支えて下さっている遠藤副委員長（富士フイルムビジネスイノベーション）、東内 WG 長（リコーテクノリサーチ）および WG メンバー各位にも改めて感謝を申し上げたい。

参考文献

1 ニューラル機械翻訳とも呼ばれている。
2 https://www.jpo.go.jp/support/j_platpat/kaizen20200518.html
3 https://japio.or.jp/service/service05_08.html
4 https://chizai.jbmia.or.jp/chi_kogyo_kikaihonyaku.html
5 https://tech-jpn.jp/tokkyo-writing-manual/
6 https://www.jbmia.or.jp/whatsnew/detail.php?id=1255
7 https://www.nict.go.jp/info/topics/2019/12/09-1.html
8 https://www2.nict.go.jp/astrec-att/member/mutiyama/pdf/2021-patent-sympo.pdf