ニューラル機械翻訳に対する特許庁UTX用語データによる用語ポストエディット

Terminological post-editing of neural machine translation results using the Japan Patent Office UTX glossary data

秋桜舎　代表
山本　ゆうじ

筑波大学を経てシカゴ大学修士号。企業向けに、大規模翻訳･文書管理／作成、日本語作文、英語の講習やコンサルを行う。近著に『IT 時代の実務日本語スタイルブック――書きやすく、読みやすい電子文書の作文技法』（2012、ベレ出版）。

http://cosmoshouse.com/（連絡用フォームから）

1　はじめに

本稿では、ニューラル機械翻訳に対する特許庁 UTX 用語データによる用語ポストエディット手法について述べる。

ニューラル機械翻訳の精度の高さ、特に流暢さについてはすでに知られているが、実務翻訳、特に体系的翻訳の観点からは、まだ多くの課題がある。特に用語に関しては、以下の 4 点の問題がある。

- 低頻度の用語の誤訳が多い
- 単語の欠落や不要な繰り返しが発生する
- 既存の用語データを反映することができない
- 用語の一貫性を保証できない

これらの問題は、ニューラル機械翻訳そのものの改善を待たずとも、適切な用語管理を行った用語データにより訳文をチェックすることで対処できる。用語チェックの過程はそのままポスト エディット工程につなげることもできる。汎用性の高い UTX 用語データをさまざまなツールで用語チェックに活用することで、ニューラル機械翻訳の用語の弱点を効果的に補うことができる。

本稿では、特許文書の Google ニューラル機械翻訳の翻訳結果に対して、3 つの翻訳支援ツール・用語ツールで、特許庁 UTX 用語データについてのチェックを行い、この手法の有効性を確認する。併せて、ISO ウィーン国際会議、産業日本語研究会文書作成支援分科会での筆者の活動についても紹介する。

2　特許庁 UTX 用語データ

特許庁では、約 13 万項目の用語データを UPF 形式（翻訳ソフト辞書形式）で公開しており、毎年 5000 語のデータが追加される。AAMT（アジア太平洋機械翻訳協会）では、特許庁の許諾を得て、この用語データを、UTX 用語集形式に変換して無償で公開した〈http://aamt.info/japanese/utx/jpo-utx.htm〉。UTX 形 式に変換することにより、各種翻訳ソフト、Excel、ISO 規格 TBX 用語集形式など、さまざまな形式に変換して活用できる。

元データの特徴として、カタカナ、ひらがな、漢字などの異表記がカバーされている用語が多数ある。また長所としては、およそ特許で扱われるあらゆる分野が含まれている。

元データの仕様については〈https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyo_dictionary.htm〉を参照されたい。

分野情報が用語に付与されていないが、植物、動物、人、組織など意味素性の情報があるため、大まかな分野情報として代用できる。また旧字体表記の項目が多数存在する。

以下に本データに含まれる用語の例を示す【表 1】。

次に、特許文書を日英のニューラル機械翻訳した結果を、この特許庁用語データを使用して翻訳支援ツールでチェックする検証の手法について説明する。

表 1　特許庁 UTX 用語データの例

分野（正確には意味素性）	項目数	例	
植物（通称、品種名、学名など）	5484	白いぼキュウリ	white spine cucumber
		メラレウカ・アルテルフオリア	Melaleuca Alternifolia
		いらくさ科植物	urticaceous plant
動物（通称、学名など）	3009	ヤブカ	striped mosquito
		モンシロチョウ	Pieris rapae
		ユーグレナ	Euglena
人（人名、職位など）	1284	昌聰	Yoshiaki
		調香士	perfumer
		登壇者	presenter
企業・組織名	4454	日本醸造協会	Brewing Society of Japan
		猟友会	hunters' association
		インド技術研究所	Indian Institute of Technology
その他	44875	化学、医学、機械、工学、その他専門用語	
分類情報なし	64726	オキシジフタル酸二無水物	oxydiphthalic dianhydride
		ガス不透性プレート	gas-impermeable plate
		パッシェン曲線	Paschen curve

3　検証の手法

本検証では、国際特許分類（IPC：International Patent Classification）の A から H までの各セクションに属する文書の要約および「請求の範囲」を Google 機械翻訳（ニューラル）で日英翻訳した。

（以下のセクション日本語名は特許庁の翻訳に基づく）

A　生活必需品
B　処理操作、運輸
C　化学、冶金
D　繊維、紙
E　固定構造物
F　機械工学、照明、加熱、武器、爆破
G　物理学
H　電気

文書の選択方法としては、7 桁の実数を乱数として生成して、特許情報プラットフォームで特許の登録番号を検索し、原文として適切な長さの（すなわち、短すぎない）文書を選択した。

ただし分野 D「繊維、紙」では、上記の方法で探すのが困難であったため、あらかじめ分野 D を特定した上で適切な長さを持つ文書を選択した。

特許文書は、日常文書と比較すると一文が極めて長く、さらにその文を慣習的に改行（ハードリターン）で区切る。これは構造的文書の観点からすると中途半端な構造であり、厄介な特徴である。箇条書きが適切な場合でも一文とされる。

長すぎる一文は、ほぼあらゆる方式の機械翻訳の精度を下げる。しかし、ハードリターンを保持するとその箇所で一文が複数に分節化され、構文が崩れてしまうため、一長一短となる。そのため、本検証は、ハードリターンを保持した文章と、削除した文章の二通りで実施した。結果的には、ハードリターンを保持したほうが人間によるチェックの観点からは有利であることが確認された。一文が長いと、原文と訳文の対応関係を見定めることが非常に困難になる。実際の考察でも、ハードリターンを保持した結果のほうを使用した。

次に、129646 項目（2017 年現在）の特許庁 UTX 用語データを翻訳支援ツール SDL Trados Studio、Memsource、および用語ツール ApSIC Xbench で使用して、翻訳結果 391 分節に対して用語チェックを行った。これらのツールのうち、Memsource と ApSIC Xbench は無償で利用できる。

4　検証結果

以下に、SDL Trados Studio、Memsource、ApSIC Xbench での用語検証の結果を示す。なお、これらの用語検証は、あくまでも用語に問題がある可能性を指摘するにとどまる。不適切な誤検出もあるので、検出語が

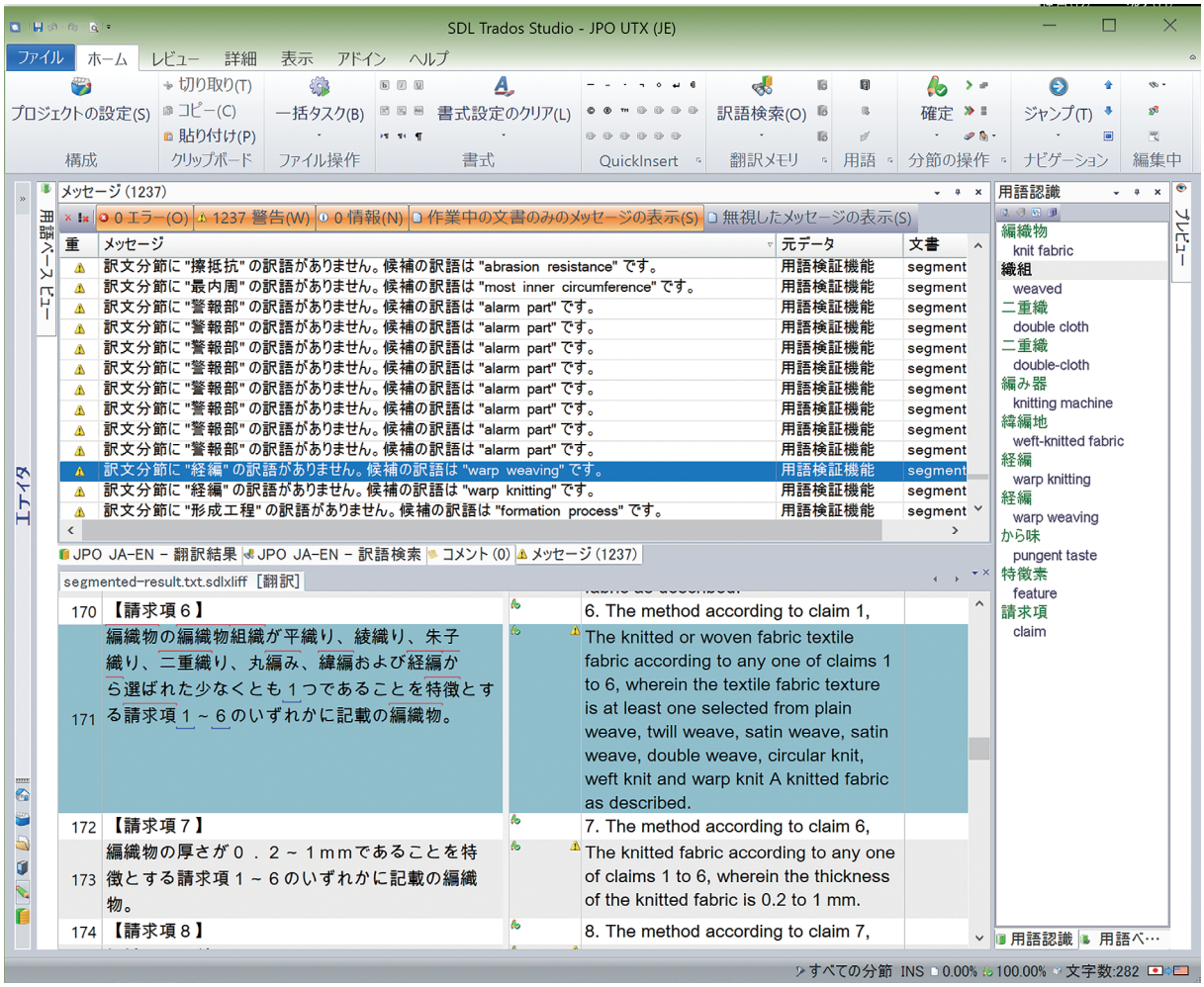

図 1　SDL Trados Studio での用語検証

多ければよいとは必ずしも限らない。

翻訳支援ツール SDL Trados Studio の用語検証機能【図 1】では、391 分節に対して 1237 件の用語検証の問題を検出した。

以下は、機械翻訳で正しく訳されていないことが用語検証で指摘された専門用語の一例である

請求項/claim

デシテックス/decitex

経編/warp weaving

センサシステム/sensor system

また、たとえば「フォージャサイト/faujasite」「焼結/sintering」という用語については、Google 翻訳は正しく訳していたが、SDL Trados Studio に用語として認識された。間違いのみではなく、専門用語が正しく訳されていることを確認できるのも翻訳者にとっては有用である。

なお原文では「挟圧して」となっていた場合に、「挟圧/compression」という語が訳されていない、という指摘がされた。実際には、「挟圧/compression」という語が名詞として、「挟圧する/compress」が動詞として登録されていた。そのため、上記の指摘は正しくない。正確には「挟圧する/compress」が訳されていない、とすべきである。とはいえ、このようなケースは、品詞解析が行われない用語チェック方式では回避が困難である。回避策としては、UTX をルールベース機械翻訳のデータとして登録し、その結果をニューラル機械翻訳と比較することで動詞の活用形の問題は検出できる可能性がある。

Memsource の用語検証【図 2】では、391 分節に対して 372 件の用語検証の問題を検出した。Trados と比較して検出した問題の数が少ない理由は、Trados ではあいまい検索が機能して、厳密な一致でない場合でも誤りの可能性を指摘するからである。

Memsource で的確に指摘できた例は、「プラネタリキャリア」という語に対する訳語が訳文から欠落する訳抜けなどである。

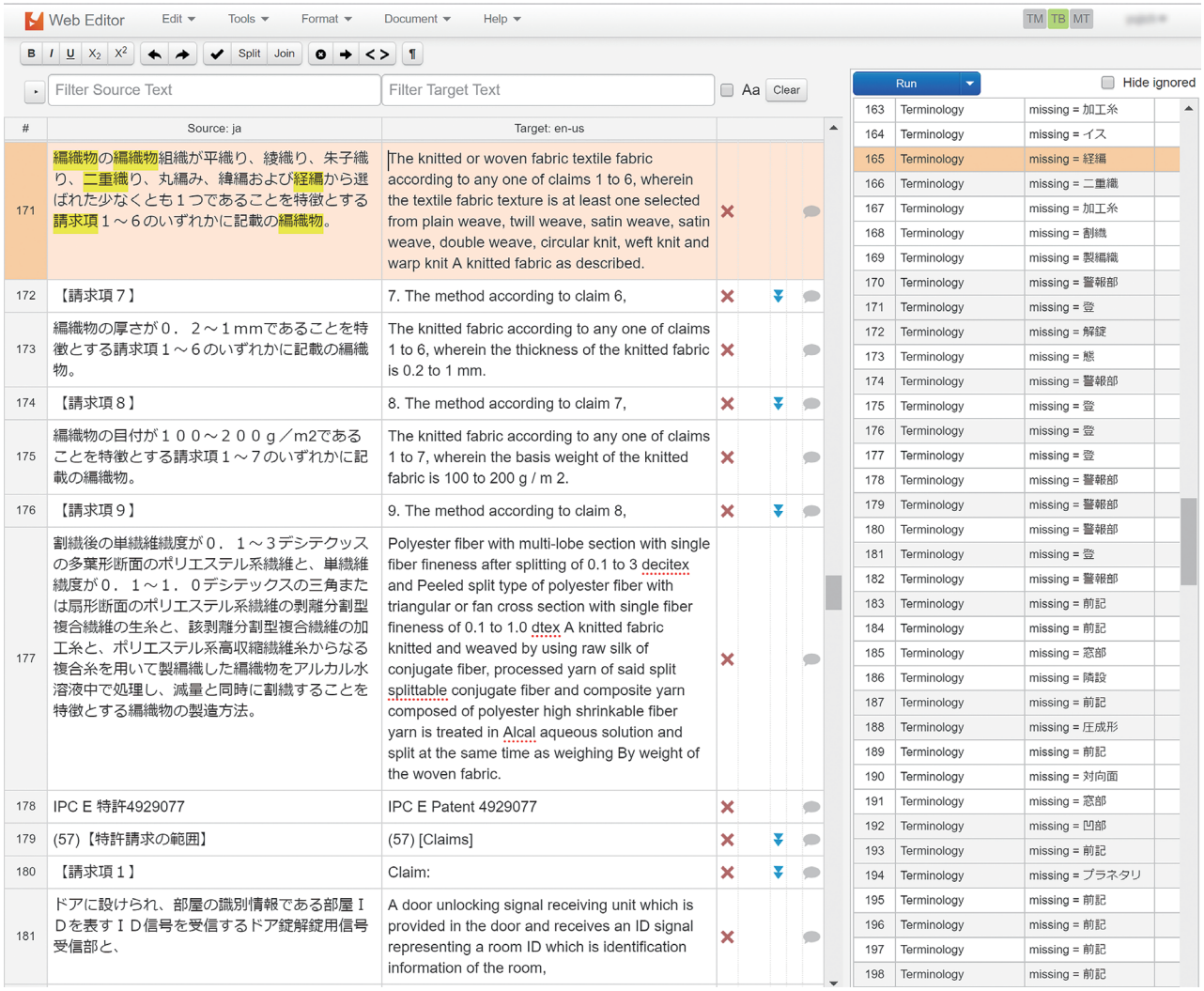

図２　Memsource での用語検証

特許庁用語データでは、「ラミ」、「マトリック」、「フィル」という特殊な用語は登録されているのに、それらを文字列の中に含む「セラミック」、「マトリックス」、「フィルター」のような用語は登録されていない。そのため、誤指摘となっていた箇所がある。これは後述するこの用語データの由来を考えるとやむを得ない。

用語ツール ApSIC Xbench の用語検証【図 3】では、391 分節に対して 603 件の用語検証の問題を検出した。結果は、Memsource と同様に厳密な用語一致のみとなるが、問題のある訳語を検出することができた。Xbench は用語チェックを行えるものの、編集環境は備えていないため、問題をみつけた場合、修正するには外部のエディター プログラムを必要とする。

３種のツールに共通の事象として、用語「前記～」（例：前記加工糸）の訳語がないという指摘が半数近くあった。（なお「前記」に対応する訳語は、特許庁用語データでは "above-mentioned" か "aforementioned" となっている。）

特に "the said processed yarn" ではなく、"said" なしに "the processed yarn " として訳された場合などである。「前記」を必ず "said" などと訳すと読みにくくなる（異論もあるだろうが）。このような指摘が不要であれば無視（あるいは用語データから削除）すればよいので、大きな問題ではない。場合によっては、「前記」が訳で抜けているかを確認したい需要もあるだろうから、有用と考えられる。

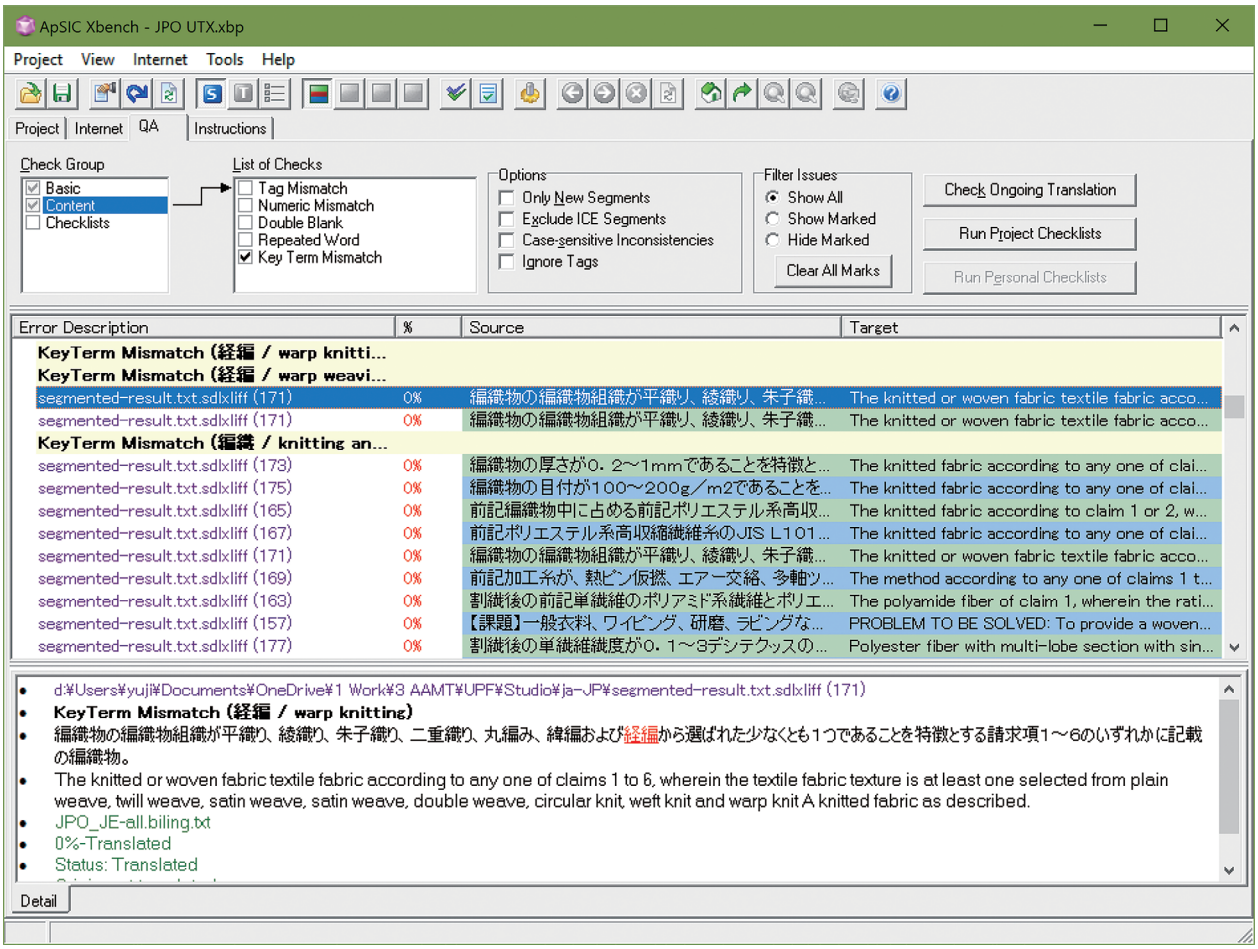

図３　ApSIC Xbench での用語検証

5 考察

特許庁 UTX 用語データと翻訳対象である特許文書の適合性は理想的と言えるほどではない。だが一定の関連性があり、改善の余地はあるものの、特許庁 UTX 用語データでの用語チェックには有用性が認められた。軽度のポストエディットのみを行うライト ポストエディットであれば、このような用語チェックだけにしぼって行う、という方法もある。これにより、ニューラル機械翻訳の弱点を効果的に補うことができる。もちろんこの手法は、フル ポストエディットにも活用できる。

特許庁 UTX 用語データを用語チェックに使用する際の課題として判明したことがいくつかある。特許庁用語データが作成された当初の目的は、ルールベース機械翻訳で未知語として処理できなかった専門用語を補うことである。そのため、「用語チェックに使うため」の調整は行われていない。だが用語チェックに使うための調整を行うことで、さらに有用性を増すことができる。具体的には、以下が可能である

1. 用語ステータスの設定
2. 不適切な用語の除去
3. 分野分け

用語ステータスの設定では、複数訳語の優先順位などの情報を付与する。本用語データには、複数訳語の優先順位の情報がない。つまり、どの訳語が最も適切かという情報がない。

たとえば「仔ヒツジ/lamb」、「仔ヒツジ/hogling」という用語がある。hogling はイギリス方言であり、一般的な訳語ではないので、特殊な場合以外には使用すべきではない。だが、それを示す情報は特許庁用語データに含まれていない。

この場合、UTX では以下のように用語ステータスを付けることができる。（以下は例示であり完全な UTX 形式ではない）

term：ja	term：en	term status：ja	term status：en
仔ヒツジ	lamb	approved	approved
仔ヒツジ	hogling	approved	non-standard

上記の表現は以下のことを示す。

- 「仔ヒツジ/lamb」の仔ヒツジと lamb の両方が approved（承認語）であること。
- 「仔ヒツジ/hogling」の hogling が non-standard（非標準語）であること。

承認語のほうが通常使用される語である。場合によっては非標準語が使用されることもあるが、例外的である。

また日本語の異表記についても同様の用語ステータスが必要になる。

多対多の用語の組み合わせとしては、深せん、深セン、深ちぇん、深チェンと Shenzhen、SHENZHEN、S. Z. があった。股分、股ふん、股フン、股フェンと Co., Ltd、CO., LTD. があった（「股分有限公司」は中国語で株式会社を指す）。

これらは企業目の固有名詞として正式表記であるなら派生型もそのまま登録するべきだが、実際のところは正式表記ではない（あるいは正式表記がどれであるか判別できない）。そのため、以下のような組み合わせの登録が多数されており、項目数が増える原因となっている。

弘凱光電（深せん）股ふん有限公司/
弘凱光電（深せん）股フェン有限公司
弘凱光電（深せん）股フン有限公司
弘凱光電（深せん）股分有限公司

上記の語の英訳語はすべて Brightek Optoelectronics (Shenzhen) Co., Ltd. である。この企業名の正式な表記が判明しているのであれば、その項目に承認語用語ステータスを付け、それ以外の項目には禁止語用語ステータスを付けることができる。こうすれば、どれが正式表記で、どれが誤記かを明確に区別できる。

用語ステータスの詳細については UTX 仕様書に記載されているので参照されたい。

〈http://www.aamt.info/japanese/utx/utx1.20-specification-j.pdf〉

また本用語データには、「態/voice」や「登/Noboru」（個人名）といった一字の用語が含まれている。さらに専門用語ではない「領域/field」のような一般語も含まれている。これらは大量の誤指摘の原因となるため、用語チェックに使用する際には用語データから除外したほうがよい。

ルールベース機械翻訳では、用語を追加すれば必ず精度が向上するというわけではない。不適切な用語が追加されれば翻訳精度は低下する。どの語を追加すべきかは慎重に吟味し、追加した後で検証が欠かせない。

特許の自動翻訳の精度を上げるには、長い一文で書く習慣を禁止することも重要だが、使用できる語彙を整理し、制限するのも効果的である。現状は「呈出する/present」、「深厚な/deep」、「尊崇する/respect」などの難解な言葉が制限なしに使われている。これらは「示す」、「深い」、「尊敬する」という語に置き換えて差し支えあるまい。上記のように、対応する英単語は複雑な語ではない。日本語だけがことさらに難解な語を使っている。

また、Trados の用語検証機能にはいくつかの課題がある。たとえば「生糸/raw silk」、「ポリアミド系繊維/polyamide fiber」、「平均厚み/average thickness」、「補正データ/correction data」が正しく訳されていたにもかかわらず、訳されていないと誤指摘された。また、複数の用語ベースを検証に使えない。これが可能になれば、分野分けした用語データの中から関連性が高いものだけを選択して組み合わせて適用できる。

この他、クラウド型翻訳支援ツールとして Matecat での用語チェックも試みたが、本ツールは用語検証機能を持っていなかった。

6 ウィーン ISO/TC37 国際会議と UTX

2017 年 6 月に開かれたウィーン ISO/TC37 国際会議について簡単に報告する（TC は technical committee（技術委員会））。筆者は ISO/TC37/SC3 の委員を務めているが、TC37/SC3/WG3（TBX グループ）関連の会議にオンラインで参加した。TBX（TermBase eXchange）は ISO で規格化された用語集形式である。

会議では、TBX について各国のコメント、TBX 仕様の厳密化および改善などがトピックとなった。TBX には、特定の用途に特化したさまざまな「方言」がある。会議では、本来の TBX とこれらの方言の関係、XML 表記スタイル、XML 名前空間などが話し合われた。TBX の重要な情報は http://www.tbxinfo.net/ で公開される。

また TBX と AAMT の関係に関しては「TC37/SC3/WG3（TBX グループ）は、同グループが、TBX

と UTX 1.20 の間の変換の定義と、資源および情報を継続的に共有することについて AAMT と協力することを推奨する」という推奨事項が採択された。

TBX と UTX は排他的なものではなく、用途、あるいは使用するステージが異なる。また相互変換できるので併用することもできる。シンプルな用語管理を UTX 形式で開始し、用語管理ノウハウが蓄積されてから TBX 形式に移行することもできる。

TBX には UTX のような用語ステータスはない。本稿で説明したように、用語ツールやルールベース機械翻訳で用語データを活用する上では、用語ステータスは重要であり、実際的である。UTX チームとしては、TBX グループに対してアピールし、今後の ISO 規格の改善に貢献できればと考える。

7 産業日本語研究会文書作成支援分科会

筆者は、産業日本語研究会文書作成支援分科会の委員を拝命しているが、この分科会では特許文書と UTX に関してオントロジー的なアプローチを提案している。具体的には、請求項を解析して、用語として抽出し、特許文書の一部を用語の観点から構造化する。【表 2】にタクシー捕捉システムの特許文書を、用語の観点からオントロジー UTX 用語データ形式として構造化した例を示す（用語データの一部は省略）。

ここでは、特許文書に含まれ、定義されている用語とそれらの上下の構成関係が「上位構成要素（x-parOf:ja）」と「下位構成要素（x-components:ja）」で示されている。このように記述することで、それぞれの用語の上位と下位の関係性が明確になり、難解な特許文書を人間にとって読みやすくでき、さらに機械的処理もしやすくなる。

なお難波ら（2014）はＦタームに基づくオントロジー構築をしている。将来的には、IPC とその日本語訳を UTX 形式とすることで、日英のオントロジー用語データを構築することも可能であろう。

8 まとめ

ニューラル機械翻訳では、用語関連の問題が指摘されているが、これは用語管理を行った用語データにより適切に補完できる。今回は、特許庁 UTX 用語データのニューラル機械翻訳結果を各種ツールでチェックすることにより、そのことを確認できた。また、汎用性の高い UTX 用語データをさまざまなツールで用語チェックに活用することで、ニューラル機械翻訳の用語の弱点を効果的に補うことが示された。

この処理では文書と用語データの適合率が高くなければならない。汎用の用語データでは適合率は低いので、文書に対して専用の用語データを作成する必要がある。ただし、特定企業内での翻訳では、一度用語データを作成すると将来的に繰り返し再利用できる。

用語管理はしばしば敬遠されているが、今後は、機械翻訳の評価方法でも、用語データを活用した評価がされることが望ましいと考える。

表 2　特許文書のオントロジー UTX 用語データ形式の例（タクシー捕捉システム）

#UTX 1.20; lang: ja/en-US; last modified date: 2017-08-23				
# 日本語	英語	品詞	上位構成要素	下位構成要素
term:ja	term:en	pos:ja	x-partOf:ja	x-components:ja
タクシー捕捉システム	taxi matching system	noun		基地局，顧客端末，車載装置
基地局	base station	noun	タクシー捕捉システム	空車位置データベース，車両データベース，地図データベース，サーバ
顧客端末	customer device	noun	タクシー捕捉システム	
車載装置	on-vehicle device	noun	タクシー捕捉システム	
空車位置データベース	available-vehicle-position database	noun	基地局	
車両データベース	vehicle database	noun	基地局	
地図データベース	map database	noun	基地局	

参考文献

難波英嗣，乾孝司，岩山真，日立製作所，櫻井孝，橋田浩一，藤井敦．“特許分類コード体系に基づくオントロジーの構築―情報分野におけるケーススタディー”．言語処理学会 第 20 回年次大会．2014.

AAMT．“UTX 1.20 仕 様 書”．AAMT. http://www.aamt.info/japanese/utx/utx1.20-specification-j.pdf,（参照 2016-8-30）.