韓国特許調査における機械翻訳クレーム検索：
機械翻訳されたクレームを検索する際の網羅性向上の検討

○田畑文也
富士フイルム（株）
〒421-0396 静岡県榛原郡吉田町川尻 4000
E-mail: fumiya.tabata@fujifilm.com

Study of the Machine translated Claim Search for Korean Patent :
Study of the Improvement of the coverage for the Machine translated Claim Search.

TABATA Fumiya
Fujifilm Corporation
4000, Kawashiri, Yoshidacho, Hibara-gun, Shizuoka , Japan
E-mail: fumiya.tabata@fujifilm.com

【発表概要】

海外特許調査において、非英語圏で出願された特許は、従来、マルチカントリー型商用 DB（データーベース）でさえクレーム以下の情報収録が無かった。近年、中国、韓国等の新興国の特許数の増大とともに DB 側の対応が進み、英語または日本語に機械翻訳された情報を収録する DB が増えてきた。また、日本特許庁より中韓文献翻訳・検索システムなども、無料で提供されるようになった。そこで、韓国特許について、商用 DB の提供する機械翻訳や、日本特許庁の中韓文献翻訳・検索システムにおいて、英語または日本語に機械翻訳されたクレームをキーワード検索する際の検索精度及び、注意点について検証した。

その結果、キーワード検索した際の再現率（得られるべき正解母集団に対する網羅性）が 9 割以上のものがある一方、2 割未満のものもあった。ヒットできなかった原因は、（1）正しいキーワードに訳されない　の他に、（2）日本語キーワードから想定されるようなハングル表記でないため、別の言葉に訳される　こともあるなどが分かった。再現率を上げるためには、（A)機械翻訳特有の異表記を考慮する　だけでなく、（B）対応するハングルの異表記から想定される機械翻訳も考慮する　などの対策が必要と考えるが、それでも機械翻訳されたものを、キーワード検索して網羅するには限界あり、現地出願人が競合として存在する場合の侵害予防調査など、網羅性が必要な場合は、漏れを防ぐため、（α）競合出願人に関しては広めに検索補完する　望ましくはさらに（β）ハングルキーワード検索により補完する　などが必要と考える。

【キーワード】
韓国特許，機械翻訳，クレーム，請求項，キーワード検索，再現率，網羅性

1. はじめに

近年、中国、韓国などの非英語圏の海外特許出願が増えており、海外特許調査において、数的には、米国、欧州以外の中国や韓国の非英語圏からの出願が 3/4 近くを占めている(図 1)。

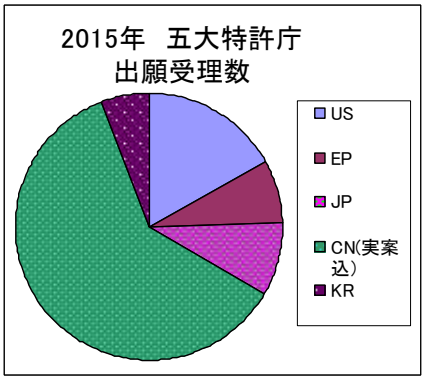

図 1. 五大特許庁への出願数

また、韓国については、技術レベル向上が顕著で、特許数が増大している(図 2)。

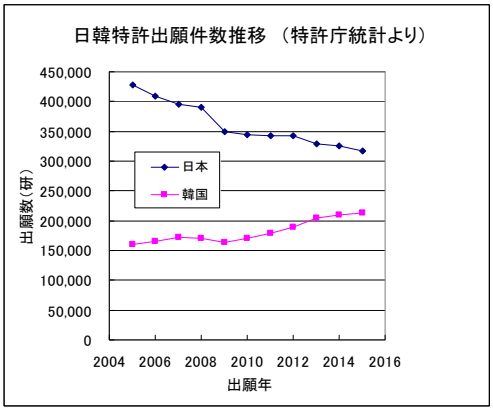

図 2. 韓国特許出願推移

特に内国人出願数の増大が顕著である(図 3)。

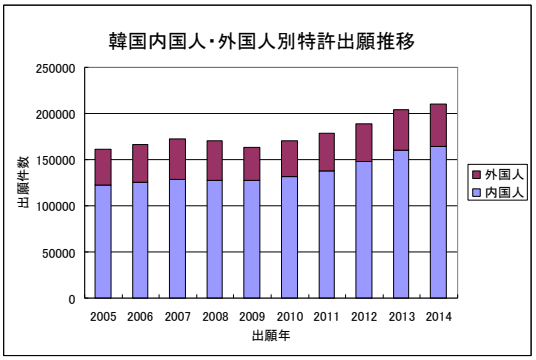

図 3. 韓国特許内国人・外国人別出願推移

韓国特許について、対応する英語または日本語で書かれた特許ファミリーがあれば、クレーム以下についてもファミリー情報を用いて、精度良くキーワード検索できるが、図 4 に示すように、5～6 割の特許が韓国のみの出願であり、これらはファミリー情報で補完することができない。

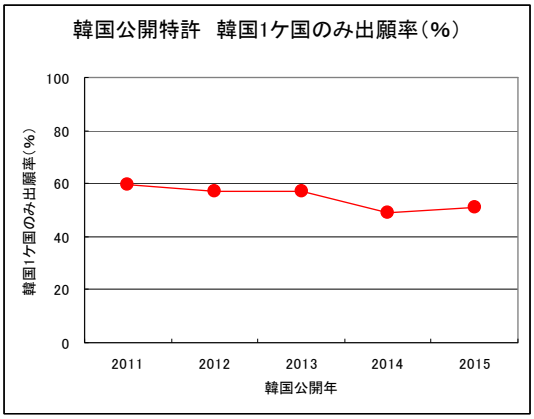

図 4. 韓国特許　韓国 1 ケ国のみ出願率
(2016 年 7 月 STN　WPI にて確認)

権利侵害防止のための遡及調査などでは、特許分類による検索だけでなく、クレームのキーワード検索は最低限行いたいところである。しかしながら、韓国特許は、ハングル(韓国語の文字)で書かれており、要約までは特許庁が人間翻訳による英語翻訳を提供するが、クレーム以下については、特許庁からは、ハングルでしか提供されない。

これらの問題に対応するため、独自に機械翻訳したクレーム以下の情報を検索できるようになっている商用 DB も増えてきており、また日本特許庁も中韓文献翻訳・検索システム(http://www.ckgs.jpo.go.jp/)を無料で提供している。これらの DB に収録されているクレームや明細書本文は、人間翻訳でなく、機械翻訳されたものであるため、これらを検索する際の注意点と、網羅性向上について検討した。

2. 検討内容

2-1） 検証に使用したキーワード

層間絶縁(膜)、ポリエチレンイミン、赤外吸収、4級アンモニウム の4種のキーワード(表 1)について、韓国 WIPS 社の PatBridge を用いて、2013 年～2015 年発行の韓国登録特許を対象に、ハングルクレーム検索した集合を用いて検証した。

表 1. 検証に用いたキーワード

キーワード概念(日本語)	対応キーワード例(ハングル)
層間絶縁(膜)	층간절연 (分かち書き有) 층간 절연 (分かち書き無し)
ポリエチレンイミン	폴리에틸렌이민 폴리에칠렌이민 (4文字目が異なる)
赤外吸収	적외선흡수 (赤外吸収) 적외선흡수 (赤外線吸収) 적외선광흡수(赤外線光吸収)
4級アンモニウム	4급암모늄(4級アンモニウム) 사급암모늄(四級アンモニウム) 4차암모늄(4次アンモニウム)

PatBridge ハングルクレーム検索式例(層間絶縁(膜)の例)

(층간절연막* or (층 adj2 간 adj2 절연 adj2 막*) or (층간 adj2 절연막*) or (층간 adj2 절연 adj2 막*)).CLA. AND (@FD>=20130101<=20151231)

2-2） 正解母集団の作成

PatBridge を用い、ハングルキーワード検索した集合をベースし、再現率を算出する際に、ファミリー特許でヒットするような影響を避けるため、トムソン・ロイター社の ThomsonInnovation にて、ファミリー特許の無いものを抽出し、かつ検索時の近接演算などで混入するノイズを目視にて除去し、正解母集団とした(図 5)。

PatBirge ハングルキーワード検索(クレーム)

↓ 特許番号移行

ThomsonInnovationでファミリーが無いものを抽出

↓ 特許番号移行

PatBridge(ハングル+日本語表示)にて、目視でノイズ除去 →正解母集団

図 5. 正解母集団の作成スキーム

なお、ノイズ除去の際は、PatBridge を用い、ハングル-機械翻訳日本語を併記し表示させ、効率的にノイズ判別した(図 6)。

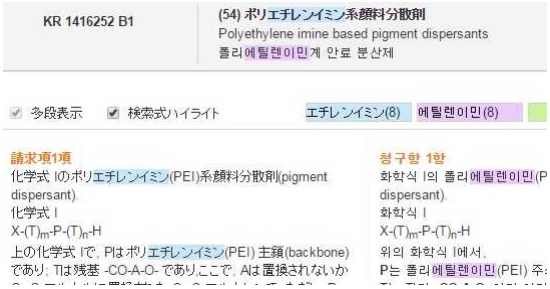

図 6. PatBridge ハングル-日本語併記表示

最終的に正解母集団として用いた公報数を表 2 に示す。

表 2. 正解母集団として用いた公報数

キーワード概念	検証正解公報数
層間絶縁(膜)	100
ポリエチレンイミン	156
赤外吸収	69
4級アンモニウム	125

2-3） 再現率算出方法

表 3 に示す DB を用い、表 4 に示すような検索式で、機械翻訳されたクレームを検索し、正解母集団との付き合わせを実施し、再現率(得られるべき正解母集団に対する網羅性)を算出した。

表 3. 機械翻訳検証に用いた DB

DB	提供元	検索に用いた言語
中韓文献翻訳・検索システム	日本特許庁	日本語
PatBridge	WIPS	日本語
ThomsonInnovation	トムソン・ロイター	英語
Shareresearch	日立	英語

表 4. 機械翻訳クレーム検索式例

キーワード概念(日本語)	日本語検索式例	英語検索式例
層間絶縁(膜)	層間絶縁 層間誘電	((interlayer*)+(interlevel*)+(intermedia*)) adj3 ((isolat*)+(insulat*)+(dielect*)) (inter adj3 (layer+level+media))*(inter adj5 ((isolat*)+(insulat*)+(dielect*)))
ポリエチレンイミン	ポリエチレンイミン ポリ adj3 エチレンイミン ポリエチレン adj3 イミン (ポリ adj3 エチレン adj3 イミン)	(Polyethyleneimine*)+PEI Poly ajd3 ethyleneimine* Polyethylene adj3 imine* Poly adj3 ethylene adj3 imine*
赤外吸収	(赤外 or IR or IR) near15 (吸収 or 吸光)	(IR+infrared) near3 ((abso*)+(filter*))
4級アンモニウム	((4級 or 4級 or 四級) adj10 アンモニウム	Quaternary* near3 ammonium*

2-4)　ヒットしなかった原因解析方法

今回の検索でヒットできなかった原因を調べるため、目視でそれらを調べ、その原因を、表 5 に示す 5 カテゴリに分け解析した。

表 5. ヒットできない原因カテゴリ定義

カテゴリ	定義
対処可能訳	予備検索を実施し、ハングル表記から、機械翻訳のくせを考慮すれば、なんとか対処可能なレベルのもの。(通常の検索を超えた対応レベル)
対処不能誤訳	機械翻訳されたものが、想像し得ない訳になっている。 機械翻訳されずに訳が消えるなど。
DB収録など	DBに収録されていない。
異義語	対応するハングル表記が、異義語を持ち、全く別の言葉として訳される。
その他	DBのインデキシングの問題など。

3. 検討結果

3-1) 再現率

4 種のキーワードについて、4 つの DB で検索した集合を、正解母集団と突き合わせて、再現率を算出した結果を図 7 (キーワード軸) 及び図 8 (DB 軸) に示す。(DB 名は明示せず)

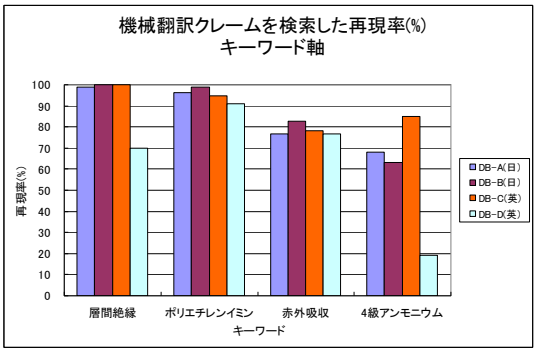

図 7.　機械翻訳クレーム検索再現率 (キーワード軸)

図 7 からわかるように、キーワード種によって大きく再現率が異なり、ポリエチレンイミンのように、4 つの DB 全てで 9 割以上の再現率のものもあるが、4 級アンモニウムのように、全体的に再現率が悪く、2 割未満の再現率しかない DB もあることがわかった。

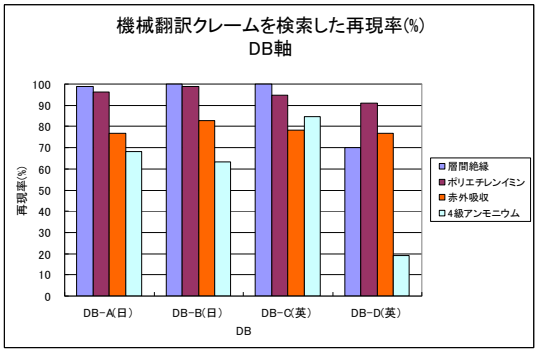

図 8.　機械翻訳クレーム検索再現率 (DB 軸)

また、DB により再現率も異なり、DB-C が比較的良い結果を示しているが、4 種のキーワードの全てについて、侵害予防調査に耐えうるレベルの再現率を有する DB は無かった。

3-2) ヒットしない原因解析

4 種のキーワードについて、日本語に機械翻訳されたもの、英語に機械翻訳されたもの分けて、そのヒットしなかった原因を図 9～図 12 に示す。

層間絶縁に関しては、日本語検索では DB の収録漏れ1件のみで、それ以外は全件ヒットした。英語検索では、層間の部分が"between a floor"と、機械翻訳されたものなどがあり、予備検索で機械翻訳の癖を掴めば対応可能なものもあるが、" getozoruyonmakku"のように、英語にはない、発音からの当て字に訳されるような対処不能誤訳が、約 7 割あった。

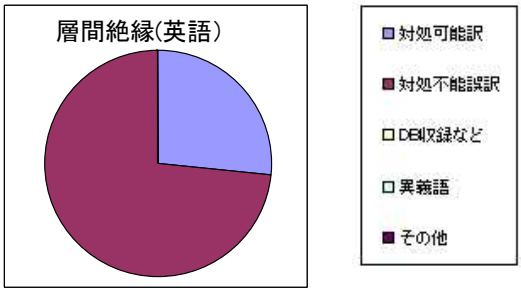

図 9.　層間絶縁　ヒットしない原因

ポリエチレンイミンについては、図 10 に示すように、日本語検索の場合、DB-A は収録に問題があり、7 割以上が収録の問題で、それ以外は”ポルリエテ ィルレンイミン“と訳される対処不能誤訳があった。英語検索では、”polyethylenemine“（イミンの”i“が抜けている）のように、ミススペルの単語に訳されたり、機械翻訳でキーワードが消失したりするなどのように、対処不能誤訳がほとんどであった。

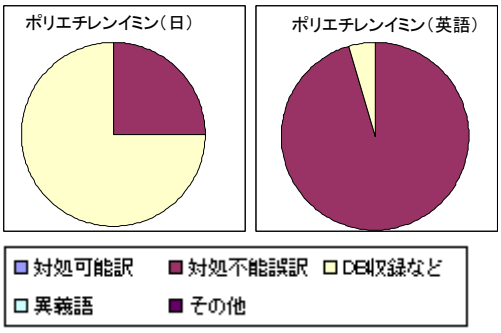

図 10.　ポリエチレンイミン
ヒットしない原因

赤外吸収について、図 11 に示すように、日本語検索では、ハングルの吸収にあたる”흡수“が、吸水という(同表記)異義語を持つため、吸水に訳されるものが、漏れのうち約 1/3 あった。また DB のインデキシングに問題あるものも漏れのうちの半数あった。英語検索においては、機械翻訳で訳が消えたり、赤外が”far-red light”と訳されたり、対処不能な誤訳をされるものが漏れの6割以上を占めた。

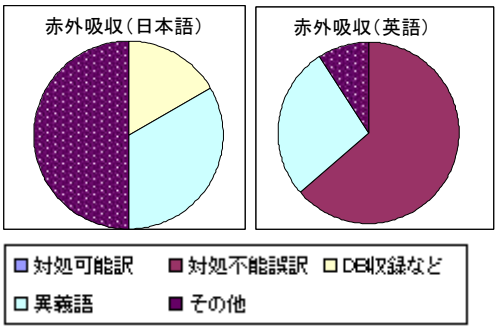

図 11.　赤外吸収　ヒットしない原因

4 級アンモニウム（quaternary ammonium）についての結果を図 12 に示す。ハングルでは、4 級アンモニウムは、大きく分けて “4 급 암모늄”(4 級アンモニウム)と、” 4 차 암모늄”(4 次アンモニウム)、つまり“級”または“次”の 2 系統で表記される。このため、機械翻訳されると、日本語では表記されないような“4 次アンモニウム”と訳されるようなものが、日本語検索における漏れのうち 7 割以上あった。同様に、英語検索では、約 9 割の誤訳が、”forth ammonium”のように訳された（対処可能訳に分類）。その他には、日本語では、“4急火アンモニウム”,” 4チァアムモニュム”など、訳されるものもあった（対処不能誤訳に分類）。

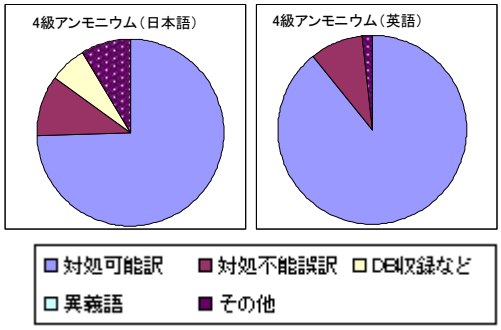

図 12.　4 級アンモニウム
ヒットしない原因

4. 考察とまとめ

韓国特許調査については、本来は、ハングルキーワード検索可能な DB で検索補完して、網羅性を上げる[1] [2]ことが望ましいいが、サーチャーの言語（ハングル）スキル、工数などの問題から、クレーム以下の情報について、機械翻訳された情報を検索できる DB で検索して対応することが実務上多い。

また、韓国特許の日本語機械翻訳検索については、以前から検討されている[3] [4] [5] が、日本語に機械翻訳されたものだけでなく、英語に機械翻訳されたも

のを含め、再現率まで踏み込んで今回検討した。

韓国特許キーワード検索において、機械翻訳されたものについても、人間翻訳されたものと同様のキーワードで検索しているがことが実務上多いが、今回の結果で、キーワード、及び DB による違いはあるものの、いずれの DB でも、4 種全てのキーワードについて、侵害予防調査に耐えるレベル再現率を有する万能的な DB はないことが判明した。すなわち、今回検討したどの DB を用いても、機械翻訳の精度には依然として問題がある。

ヒットできなかった原因を解析すると、
(1) 正しいキーワードに訳されない
の他に、
(2) 日本語キーワードから想定されるようなハングル表記でないため（例：4 級アンモニウム→4 次アンモニウムの場合有り）、別の言葉に機械翻訳されることもある
などが分かった。他にも、日本語の吸収に対応するハングルの”音仐“が、吸水と言う同表記異義語になっている場合もあり、漢字も持たないハングルでは、このような例は少なくない。（表 4）

表 4. ハングル同表記異義語例

ハングル	日本語
산	酸 山
충전	充電 充填
발전	発電 発展
진통	陣痛 鎮痛

再現率を上げる為には、
(A)機械翻訳特有の異表記を考慮するだけでなく、
(B) 対応するハングルの異表記から想定される機械翻訳も考慮するなどの対策が必要と考えるが、それでも機械翻訳部をキーワード検索して網羅するには限界あると考える。現地出願人が競合として存在する場合の侵害予防調査など、網羅性が必要な場合は、漏れを防ぐため、
（α）現地競合出願人に関しては、特許分類を含め、広めに検索補完する
望ましくはさらに
（β）ハングルキーワード検索により補完する
などが必要と考える。

5. 終わりに

最後に、本報告は 2016 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] 田畑文也 他 : "網羅性のある韓国特許調査"（第 9 回情報プロフェショナルシンポジウム、2012/10）
[2] 田畑文也 他 : "韓国特許調査手法の検討”（第 11 回情報プロフェショナルシンポジウム、2014/12）
[3] 前田佳治 他
"日本語で検索できる特許データベースの検証：（韓国特許データベースについて）"（第 5 回情報プロフェショナルシンポジウム、2008/10）
[4] 佐野浩太郎 他 : "中韓文献翻訳・検索システムの検証”（第 12 回情報プロフェショナルシンポジウム、2015/12）
[5] 西尾潤 他 : "中韓文献翻訳・検索システム：検索漏れの低減案”（第 12 回情報プロフェショナルシンポジウム、2015/12）