平成 23 年度

特許文献の機械翻訳のための辞書データ整備に関する調査

調 査 報 告 書 （概要版）

平成 24 年 2 月

特　許　庁

1. 調査目的

従来から特許出願が多くアクセスの必要性が高かった英語、ドイツ語等に加え、中国、韓国等の非欧米諸国の特許出願が近年増大している。英語文献については、我が国の審査官および出願人が原文でその発明の内容を理解することが可能であるが、非欧米言語を原文のみから正確に理解することは困難である。このため、機械翻訳を活用してこれらの文献の日本語訳を作成し、多言語で記載された特許文献を日本語で検索できる環境整備の必要性が高まっている。

しかしながら、その検索環境整備に向けた調査の結果、特に中国語についてはその翻訳の難しさから特許審査で利用できるレベルには達しておらず、翻訳精度向上のためには大幅な辞書整備が必要である。

また、既に特許庁が独立行政法人工業所有権情報・研修館の事業を通じて提供している「特許電子図書館」英語版や「高度産業財産ネットワーク」では、特許公報や特許審査関連情報を日英機械翻訳を利用して英語で提供しており、この日英翻訳精度の向上のためにも特許文献の翻訳に適した機械翻訳辞書の整備は重要である。

このような状況を踏まえ、本調査は、外部有識者の専門的知見を活用し、特に、中日、韓日機械翻訳システムおよび既存の日英機械翻訳システムにおいて、特許文献の翻訳精度向上に資する辞書データの効率的な作成方法について調査を実施したものである。

2. 調査方法

本調査では、特許文献に特化した機械翻訳辞書データ整備に向け、（1）機械翻訳辞書作成方法調査、（2）民間開発業者が提供する機械翻訳ソフトウェアに関する調査、ならびに（3）特許庁が利用できる辞書等の有効活用に関する調査を行い、特許文献の翻訳精度向上に資する辞書データの効率的な作成方法およびその有効活用について検討を行った。

その検討に際しては、外部有識者の専門的知識・経験を活用するため、委員会形式をとった。委員会の委員の構成は以下のとおりである。

委員長	中川　裕志	東京大学　情報基盤センター　教授
委　員	黒橋　禎夫	京都大学 大学院情報学研究科 知能情報学専攻　教授
	宇津呂　武仁	筑波大学 システム情報系　知能機能工学域 准教授
	隅田　英一郎	独立行政法人情報通信研究機構　多言語翻訳研究室室長
	金　楓	株式会社高電社 ソフトウェア開発室　主任
	熊野　明	東芝ソリューション株式会社　プラットフォームソリューション事業部　クラウドサービス商品技術部 参事

3. 調査結果

3.1. 機械翻訳辞書作成方法調査

パテントファミリーからの辞書作成、他言語を中間言語とした辞書作成、人手翻訳を利用した辞書作成を中心に、以下８つの辞書作成方法について調査を実施した。

（１）パテントファミリーからの辞書作成
（２）他言語を中間言語とした辞書作成
（３）人手翻訳を利用した辞書作成
（４）機械翻訳の未知語からの辞書作成
（５）コンパラブルコーパスからの辞書作成
（６）中国語と日本語の対応する漢字に注目した辞書作成
（７）韓国語と日本語の読みに注目した辞書作成
（８）Wikipedia を利用した辞書作成

3.1.1. パテントファミリーからの辞書作成

パテントファミリーからの辞書データ生成は、パテントファミリー関係にある複数の特許文献対から対訳コーパスを作成し、作成した対訳コーパスから対訳辞書を作成する２つのステップからなる辞書データ作成方法である（図 3-1)。

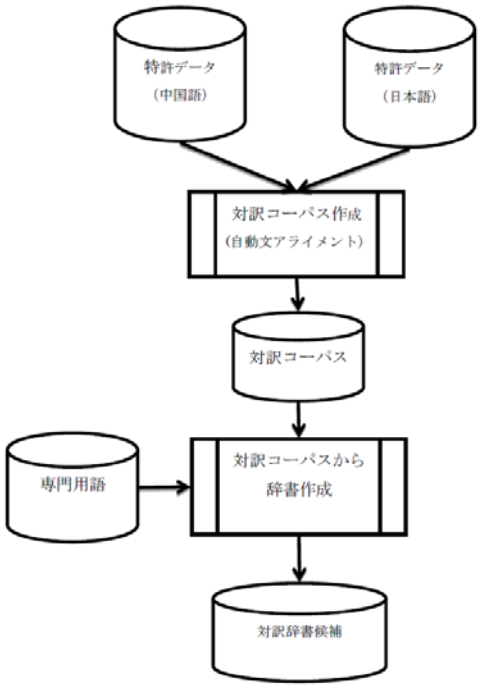

図 3-1 パテントファミリーからの辞書作成

パテントファミリーからの辞書作成では入力するデータが特許文献である。そのため、一般に購入する辞書と異なり、作成される対訳辞書は実際に特許文献内で使用された用語となる。さらに、特許文献には IPC（国際特許分類）が付与されている。このことから、IPC を利用した分野別の対訳辞書作成も期待できる。

また、この辞書作成方法の精度は作成できる対訳コーパスの精度に影響を受ける。そのため、逐次翻訳したものは、対訳コーパスとしての利用価値が高い。例えば、国際特許出願（PCT 出願）は公開言語から国内移行した国の言語への逐次翻訳が求められることから、パリルートの文献対よりも良質な対訳コーパスが取得できると考えられる。同様の観点から、先行技術文献調査等の目的のために作成されている、特許文献の要約の逐次翻訳による対訳データを利用すると、良質な対訳辞書が取得できると考えられる。

パテントファミリーからの辞書作成方法により中日辞書を作成する場合は、一定の技術的困難性はあるものの、日英等の他の言語間では既にこの方法により辞書が作成されており、その専門的知見を有する企業、大学、研究機関等の知識・経験を利用できるとすれば、技術的に可能であるといえる。また、この辞書作成方法は、大量の特許文献から得られる対訳コーパスを利用して、特許文献で使用されている用語の大規模な対訳辞書の作成が期待できることを踏まえると、辞書利用可能性や実効性は高いと考えられる。

3.1.2. 他言語を中間言語とした辞書作成

他言語を中間言語とした辞書作成は、主に英語を中間言語として２つの既存辞書を組み合わせて対訳辞書を作成する方法である。例えば、中英辞書と英日辞書を組み合わせることで中日辞書を作成することが可能であると考えられる（図 3-2）。

他言語を中間言語とした辞書作成は、中間言語とする言語を有する２つの対訳辞書が必要である。

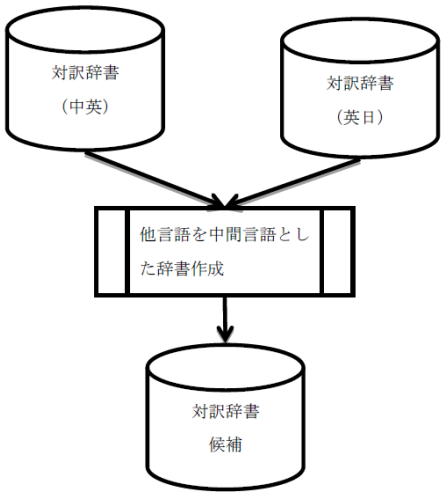

図 3-2 他言語を中間言語とした辞書作成

この辞書作成方法の技術的困難性の程度は低いものの、作成できる辞書の規模は使用できる辞書の量が上限となる。また、作成される辞書の分野は入力される２つの対訳辞書の分野に依存するため、必ずしも特許文献の翻訳に適した辞書が作成できるとは限らない。さらに、使用する辞書の利用条件についても考慮する必要がある。

3.1.3. 人手翻訳を利用した辞書作成

人手翻訳では、翻訳対象となる文献の量が多く、複数の翻訳者による翻訳作業が必要となる場合等に、翻訳結果の用語統一を目的として用語集を作成することがある。よって、人手翻訳を行う際にこのような用語集を作成し、辞書として用いることが可能である。

この辞書作成方法は、人手による作業が多くを占めるため、作成にかかる費用は他の辞書作成方法に比較して極めて高額になる。

3.1.4. 辞書作成方法の比較検討

各辞書作成方法の文献調査の結果から、作成期間、費用、作成される辞書の品質、獲得できる辞書規模を比較・検討した（図 3-3）。

作成期間、費用の観点からは、パテントファミリーからの辞書作成と、他言語を中間言語とした辞書作成が適していると考えられる。人手翻訳を利用した辞書作成に要する費用は極めて高額であるため、辞書作成のために人手翻訳を実施することは費用の面で効率的ではない。

作成される辞書の品質の観点からは、パテントファミリーからの辞書作成は実際の特許文献に出現する用語から辞書を作成するため、特許文献の翻訳に適した辞書が期待できる。一方、他言語を中間言語した辞書作成により得られる辞書は、作成元となる辞書の分野に依存するため、必ずしも特許文献の翻訳に有効とはいえない。

獲得できる辞書の規模を考えると、パテントファミリーからの辞書作成は、大量の特許文献を利用した大規模な辞書作成が可能である。

これらの点から総合的に判断すると、特許文献の機械翻訳の翻訳精度向上のために、大規模な機械翻訳辞書整備を効率よく行うには、パテントファミリーからの辞書作成方法が最も適していると考えることができる。

辞書作成方法	辞書作成方法の概要	事業実施の実現性検証項目			
		辞書整備に要する時間（20万語規模対訳辞書）	初期投資を含めた費用（20万語規模対訳辞書）	作成される辞書の品質（特性）	獲得できる対訳辞書の規模
パテントファミリから辞書作成	・パテントファミリーから対訳辞書を作成する方法 ・主に統計的手法を用いる。	3ヶ月 （機械処理4週間+人手チェック2ヶ月）	1,770万円 （機器100万円+システム開発1,100万円+人手チェック570万）	・特許文章に出現する表現が期待できる。 ・入力の対訳コーパス、日本語用語数の量に応じて大量の対訳辞書獲得が期待できる。	327万語
他言語を中間言語とした辞書作成	・主に大規模な対訳辞書が存在しない言語対の対訳辞書を作成するための方式 ・既存の対訳辞書を用いる。	2．1ヶ月 （機械処理0.1ヶ月+人手チェック2ヶ月）	1,060万円 （機器20万円+システム開発600万円シ+人手チェック440万円）+既存辞書購入費用	・用語自体の正確性は使用する辞書の品質に依存する。 ・複合語の抽出精度は、単語の抽出精度に比べ低い	21万語
人手翻訳を利用した辞書作成	・翻訳過程にて対訳用語を収集する方法	6ヶ月	2億3,634万円 （翻訳費用：1億9,734万円+ 辞書作成：3,900万円）	・人手による作成のため品質は高い ・人手作業に係る費用が高額になる。	91万語

図 3-3 辞書作成方法の比較

3.2. 民間開発業者が提供する機械翻訳ソフトウェアに関する調査

3.2.1. 機械翻訳ソフトウェア調査

中日、韓日、日英翻訳ごとに、主要な民間機械翻訳開発業者 5 社が提供する機械翻訳ソフトウェア等について、ルールベース方式等の翻訳方式、入力可能な辞書データの形式、製品の基本辞書・専門用語辞書の収録語数等を調査した。

調査結果から中日、韓日機械翻訳ソフトウェアが備える辞書用語数は、日英辞書用語数に比べ十分ではないことがわかった（表 3-4)。

中日機械翻訳辞書用語数

項目	製品A	製品B	製品C	製品D	製品E
基本辞書	300,000	300,000	260,000	64,000	300,000
専門用語辞書	157,545	287,000	310,000	40,500	381,000
（用語数,分野数）	20	13	回答なし	5	11

韓日機械翻訳辞書用語数

項目	製品A	製品B	製品C	製品D	製品E
基本辞書	300,000	回答なし	200,000	260,000	260,000
専門用語辞書	0	0	0	575,000	0
（用語数,分野数）	0	0	0	14	0

日英機械翻訳辞書用語数

項目	製品A	製品B	製品C	製品D	製品E
基本辞書	1,270,000	1,050,000	1,430,000	1,060,000	2,010,000
専門用語辞書	627,687	3,137,000	2,759,000	3,255,000	1,505,000
（用語数,分野数）	20	37	28	48	37

表 3-4 機械翻訳ソフトウェアの辞書用語数

3.2.2.　3.2.2　辞書データ形式及び汎用性の調査

3.2.2.1. 辞書形式の規格

特許庁が今後作成する辞書データの市販機械翻訳エンジンへの適用可能性を検討するため、UTX 形式等の機械翻訳辞書データのデータ形式の規格を調査した。

辞書形式の規格として、アジア太平洋機械翻訳協会（AAMT）で策定された辞書フォーマットである UPF 形式、UTX 形式の２形式が存在した。特許庁が作成する辞書形式は、中日・韓日辞書については、中国語や韓国語を表現できるように文字コードを UTF8 とした UPF 形式、もしくは UTX 形式を採用することができる。また、日英については、UPF 形式もしくは UTX 形式のいずれも採用することができる。

3.2.2.2. 辞書作成対象とする品詞

辞書作成方法の汎用性や実現性等の観点から、特許庁が作成する辞書の項目は見出し語、訳語、品詞（見出し語)、品詞（訳語）とし、辞書作成対象とする品詞は名詞、および日本語でサ変動詞になるような中国語等の動詞を作成するのが適当であるとの方向性を得た。

3.3.　特許庁が利用できる辞書等の有効活用に関する調査

3.3.1.　無料又は市販の辞書データ、対訳コーパスの現状調査

無料又は市販の辞書データ、対訳コーパス、ならびにその規模、技術分野等の特性、利用条件等について調査を行った。

中日辞書に関しては、技術用語、専門用語が収録されている辞書が３つ存在した（「ライフサイエンス辞書」他)。ただし、本調査による特許技術用語のカバー率調査の結果から、これらの辞書を特許文献の機械翻訳に利用する場合、その効果は限定的であって、これらの辞書のみでは特許文献の機械翻訳の十分な品質向上は困難であると考えられる。

韓日に関して、辞書、対訳コーパスとも、特許・技術文献の翻訳に有用な言語資源を発見することができなかった。

日英については、「JST 機械翻訳辞書」、「日英英日専門用語辞書」等の辞書データが存在した。

3.3.2.　辞書データ、対訳コーパスの学習データとしての有効性調査、翻訳アルゴリズムについての調査

主な機械翻訳の方式である、ルールベース機械翻訳、統計的機械翻訳、用例ベース機械翻訳、それらを組み合わせたハイブリット翻訳において、辞書データ、対訳コーパスの学習データ等としての利用可能性について調査した。その結果、どの方式の機械翻訳であっても辞書データ、対訳コーパスは翻訳精度向上のための有用なデータとなりうるとの知見を得た。

3.3.3.　可逆性の調査

外国特許庁との間で辞書データの交換を行った場合に生じうる、辞書の可逆性の問題とその解決手段について、調査を実施した。委員会では、特許文献の機械翻訳で課題となる専門用語は、通常、複数言語間での概念のずれが少なく、多くの専門用語は可逆であるとの見解があった。調査を行った日英辞書では、85%について日本語と英語が１対１の関係であり、多くの用語の可逆性が確認できた。また、見出し語と訳語が１対ｎの場合についても、それぞれの訳語の分野ごとの利用頻度により、各訳語の分野別の適訳を判断して１対１の関係となるよう整理することで、翻訳方向の異なる辞書を利用することが可能となる。

3.3.4.　辞書作成規模調査

中日翻訳精度の向上に資する中日辞書の辞書作成規模について、委員会において、現状で翻訳精度が十分に高い英日・日英機械翻訳と同程度の品質を得るには、英日・日英翻訳と同程度の、基本用語及び専門用語の合計１００万語規模の辞書を整備することが一定の目安になるとの見解があった。

これを確認するため、日本の特許公開公報の明細書から汎用的な言語処理ツールを用いて、明細書中に出現する用語を実際に抽出し、特許庁が作成すべき辞書規模の目安を推定した。

また、この調査において、用語と出現回数の関係を調査したところ、出現回数がいわゆるロングテールであることが確かめられた（図 3-5）。出現回数が少ない語の辞書を作成することは非効率と考えられるので、出現回数が一定回数以上（例えば、少なくとも５以上）の用語の辞書を作成することにより、効率よく辞書作成が可能となる。

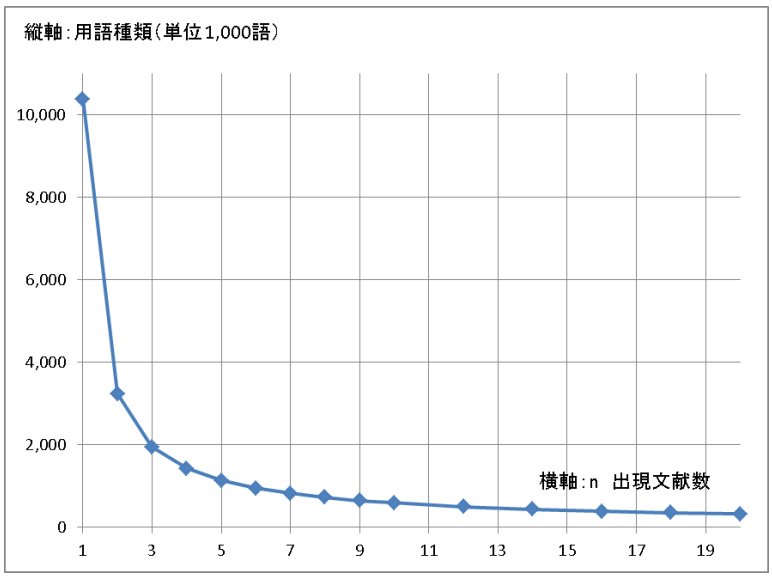

図 3-5　出現回数ｎ以上の用語の出現文献数（26 万文献の場合）