特許文の訳し分けにおける格フレームの有効性*

鈴木　勘平　　横山　晶一
山形大学大学院理工学研究科

1. 研究概要

近年、特許のような知的財産が社会において重要な存在として認識されてきている。これに伴い、特許申請数及び国際化における国際特許が増加傾向にあり、正確で迅速な機械翻訳が求められている。日英機械翻訳における訳文品質の分析[1]において、訳文品質低下の最大の原因は訳し分けの不適切さであると報告されている。訳し分けの精度を向上させるためには、文中で使用された単語の意味(語義)を解析する必要がある。

本研究では、単語の意味を特定する手掛かりとして格フレームを利用し、日英翻訳された特許文を用いてその有効性を調査した。

2. 背景

2.1 特許文

特許文は他の特許との差異を明確にするために長くて複雑な文章になりがちである。本研究では、資料としてJapio研究会特許情報データベース[2]を用いる。これは特許公報をデータベース化したものである。【要約】の部分は日本語訳と人手で翻訳された英語訳が収録されており、この対訳を研究に使用する。

2.2 格フレーム

格フレームとは、動詞を基準として取り得る格とその値に関する制約を記述したものである。本研究では、KNP[3]の格解析結果として出力される格フレーム[4]を利用する。この格フレームはWeb上の約16億文の日本語テキストから自動的に構築しており、約4万用言から成っている。例として、動詞「積む」の《動1》及び《動3》の格フレームを図1に示す。

訳し分けにおける格フレーム利用の具体例を示すために以下の例文を挙げる。

(a) {選手}が{経験}を　**積む**。
(b) {従業員}が{荷物}を　**積む**。

積む/つむ:動 1
<ガ格> 選手:17,自分:14,人:10,...
*<ヲ格> 経験:37342,体験:1363,...
<ニ格> <補文>:70,実際:36,それぞれ:8,...
<デ格> 現場:121,会社:96,分野:86,...
積む/つむ:動 3
<ガ格> 人:12,職人:6,男:6,...
*<ヲ格> 荷物:3592,石:1074,荷:884,...
<ニ格> 車:739,トラック:100,船:77,...
<デ格> 河原:24,手:6,前:5,...

図1「積む」の格フレームの一部

それぞれの文をKNPに入力し格解析を行うと、例文(a)の「積む」は《動1》、例文(b)の「積む」は《動3》となり、“異なる格構造を持つ”ということがわかる。

また、動詞を格フレーム毎に英訳すると、《動1》の「積む」は[acquire](蓄積する)、《動3》の「積む」は[load](荷を載せる)となるため、例文(a)の「積む」には[acquire]、例文(b)の「積む」には[load]が適切であることがわかる。このように、“格構造の違い”を“英訳の違い”と捉えることができるかを検証する。

2.3 先行研究

「格フレームを用いた特許文の訳し分け」[5]において、特許文を対象に、格フレームの違いが英語の訳し分けに対応するのかどうかが検討された。結果としては「格フレームは役立つ可能性がある」という段階に止まっていたが、解析作業の非効率性や格フレーム辞書自体の不備が問題点として挙げられていた。

また、新聞データの日英翻訳を対象とした研究である「結合価文法による動詞の訳語選択能力の評価」[6]において、結合価文法を用いたことによって訳し分け精度の向上が見られたことが報告されている。

3. 研究内容

3.1 使用するデータ

本研究では、日本語で書かれた特許文をKNPに入力して構文解析を行い、格フレームの解析結果を参照する。KNPへの入力文は特許文全文

*Efficiency of Case Frames in Translation Disambiguation of Patent Sentences by SUZUKI Kanpei & YOKOYAMA Shoichi, Yamagata University

の中から 2004 年度に公開された C12N 分野の【要約】の項について抜き出した 6251 文を使用する。

3.2 実験1:格フレームと英訳の対応

入力文 6251 文の中から、出現頻度の高い動詞「結合(する)」、「制御(する)」、「置換(する)」、「含む」、「示す」、「用いる」が含まれる文をそれぞれ抜き出し KNP により格解析を行い、英訳との対応を調査した。

動詞「含む」の解析結果の上位 3 つの格番号とそれに対応する英訳を表 1 に示す。入力文からランダムに抜き出した 100 文について解析を行った。「含む」の英訳は 4 種類あったが、表 1 のように一つの格フレームに対して複数の英訳に翻訳され、“格構造の違い”と“英訳の違い”は一対一に対応していないという結果となった。また、他の動詞に対しても同様の結果が得られた。

3.3 実験2:英訳と日本語文の格構造の対応

次に、対応する英訳の違いが日本語の動詞の意味を分ける手掛かりになるという考えの下、ある動詞について、英訳毎に日本語文を収集し格構造の傾向を調査した。

動詞「含む」についての解析結果を図 2 に示す。図 2 より、取り得る格については大きな違いは見つからなかったが、共起する名詞である格要素に注目すると、[contain]は格要素として具体物を取り、[comprise]は抽象物を取るという明確な違いが見られた。しかし、[include]についてはどちらの格要素も出現するという結果になった。また、[involve]と訳された日本語文は 1 文のみであった。このことから、[contain]と[comprise]については、用法の違いを表現する格構造を構築できることが分かった。

表 1 「含む」の解析結果

格番号	英訳	英訳の数	
動 3	contain	41	67
	comprise	16	
	include	10	
動 1	contain	9	10
	include	1	
動 7	contain	5	9
	comprise	3	
	include	1	

contain(65) <ガ格>培地:6、蛋白質:5、DNA:5、... <ヲ格>配列:8、ポリペプチド:7、DNA:6、... <ニ格>塩基配列:1
comprise(19) <ガ格>方法:13、試薬:2、液滴:2、... <ヲ格>工程:5、こと:5、緩衝液:2 ...
include(15) <ガ格>方法:3、DNA:2、疾患:1... <ヲ格>工程:3、遺伝子断片:1、癌:1 ... <ニ格>下流:1

図 2 「含む」の各英訳に対する格構造

4. まとめ

本研究では、Web から自動構築された格フレームが特許文の訳し分けに有効であるかどうかを調査した。結果として、一つの動詞に対する格フレーム数が多く、その全てが必ずしも用法毎に整理されていないという点から訳し分けへの適用は困難であった。そこで英訳毎に日本語文を収集し格構造を調査したところ、格要素を動詞の用法毎に分類でき、格フレームのような構造を構築できる可能性があることが分かった。

今後はさらに他の分野での調査を進める。また、作業効率向上のためにある程度の自動化を考える必要がある。最終的には既存の格フレームを組み合わせることで、より動詞の意味を捉えた格フレームの構築が期待できる。

謝辞

特許データベースを提供して頂いた AAMT/Japio 研究会に感謝致します。

参考文献

[1]麻野間直樹, 中岩浩巳: 目的言語の単語共起情報を利用した訳語選択と未知語の訳出, 言語処理学会第5回年次大会論文集, pp.442-448, (1999)

[2](財)日本特許申請情報機構:AAMT/Japio 研究会特許情報データベース(2004)

[3]KNP: 京都大学黒橋研究室(http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html)

[4] 河原大輔, 黒橋禎夫: 高性能計算環境を用いたWebからの大規模格フレーム構築, 情報処理学会 自然言語処理研究会 2006-NL-171, pp.67-73(2006)

[5]奥山真澄,横山晶一:格フレームを用いた特許文の訳し分け, 情報処理学会東北支部 2008 年度第 6 回研究会 (2009) B-1-3

[6]金出地真人, 徳久雅人, 池原悟, 村上仁一: 結合価文法による動詞の訳語選択能力の評価, 自然言語処理, Vol.11, No.3, pp149-164(2004)