語のグループ化を用いた特許文動詞の自動訳し分けに関する調査

山形大学大学院理工学研究科教授 横山 晶一

PROFILE

1949 年生。1972 年東京大学工学部卒。同年電子技術総合研究所入所。1991 年同所知能情報部自然言語研究室長。1993 年4月より山形大学。現在大学院理工学研究科教授（情報科学分野）。工学博士。アジア太平洋機械翻訳協会 (AAMT)Japio 特許翻訳研究会副委員長。

yokoyama@yz.yamagata-u.ac.jp TEL 0238-26-3336

山形大学大学院理工学研究科専攻 高野 雄一

PROFILE

1986 年生。山形大学大学院博士前期課程情報科学専攻２年

1 はじめに

特許文が長くて複雑な構造を持つことは、すでに周知の事実であり、これについては何度も言及してきた [1 ～ 3]。そのような複雑な構造を持つ特許文を機械翻訳する場合、訳文の品質は、当然のことながらそれほどは期待できない。

もともと、日英機械翻訳における訳文品質の分析 [4] において、訳文品質低下の原因は訳し分けの不適切さであると報告されている。訳し分けとは、ある文中の単語を翻訳するときに訳の候補が複数ある場合、その文に最も適した訳を選択するということである。例えば、「当てる」という動詞は、「バットにボールを当てる。」という文では「hit」、「壁に手を当てる。」という文では「put」と訳される。この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。

本稿では、単語の意味解釈をした上での訳し分けのために、文章を意味のつながりで示すことの可能な特許文独自の語のグループ化を行う試みについて報告する。語のグループ化とは、従来から行われているように、“「男」、「少女」→<人>”、“「荷物」、「鞄」→<具体物>”など、語を分類する方法のことである。語のグループ化が訳し分けに役立つかどうかを調べ、従来よりも精度の高い訳し分けが可能なシステムを作ることを目標とする。

なお、本稿は、主として [5] に基づくものである。

2 関連研究

特許文訳し分けの研究として、鈴木 [6] は格フレームと英訳との対応について調査した。結果として、「含む」に対応するどの英訳についても取りうる格に違いは見られなかった。しかし、[contain] は格要素として具体物を取り、[comprise] は格要素として抽象物を取るという単語間での明確な違いが見られた。そのため、動詞「含む」の意味の違いをある程度表現できる可能性があると結論付けている。

網川ら [7] は訳語選択の知識をコンパラブルコーパスから獲得する手法を提案している。また、単言語における語義曖昧性解消は、辞書やコーパス等のデータを使って教師なし学習を行う手法が提案されてきた [8]。この文献では、品詞等の文法的情報、構文的関係にある語、および周辺に共起する分野に関する語を曖昧性解消の手掛かりとして用いている。Dagan and Itai[9] は翻訳先言語の単言語コーパスを用いた語義曖昧性解消手法を提案している。Li and Li[10] は単語の翻訳の曖昧性解消を対訳辞書の関連付けを元にブートストラッピングによって分類器を構築し行っている。Vickrey et al [11]

は統計的機械翻訳に文脈を考慮した訳語選択を素性として導入し、訳語の選択を試みている。文全体の統計的機械翻訳システムへの導入を大きな課題としている。

3 調査内容

3．1．システム作成の目標

本研究では、意味解釈をした上での訳し分けのために特許文独自の意味グループ辞書を作成し、訳し分け精度を向上したシステムの作成を目指す。訳し分け辞書の作成手順を図 1 に示す。手順は以下のようになる。

①特許文独自の意味グループ化辞書を作成する

②英訳の付属している学習データを述語項に分け、各文をグループ化する

③グループ化を行った文と、それに対応する英単語を訳し分け辞書に追加する

翻訳の際に訳し分け辞書を用いて、機械翻訳を行うことで意味を考慮した上での翻訳を可能とする。

3．2．調査 1

まず、意味解釈が機械翻訳の訳し分けにおいて有用であるか調査を行った。自動翻訳サイトを利用した場合の特許文訳し分け精度を調べた後、グループ分けに基づいて置換した場合の訳し分け精度を調べた。

今回使用した特許文データは 2004 年度に公開された A61B 分野の [要約] の項について抜き出した文を扱った [12]。この特許文には日本語文とそれを人手で英訳した文が収録されている。今回は出現数が多く、複数の訳し分けがある動詞として「含む」([include：全体の一部として]、[contain：中に持っている]) を扱った。特許文から「含む」の対訳が [include] と [contain] となる文を無作為にそれぞれ 30 文ずつ収集した。日本語から英語への機械翻訳には Web の自動翻訳サイト「Infoseek マルチ翻訳 [13]」「Google 翻訳 [14]」「excite 翻訳 [15]」「Yahoo! 翻訳 [16]」を利用した。

特許文から抽出した文から日本語の本文を、各翻訳サイトで日英翻訳をして訳し分けをされるかどうかを調べた。以下に結果を示す。

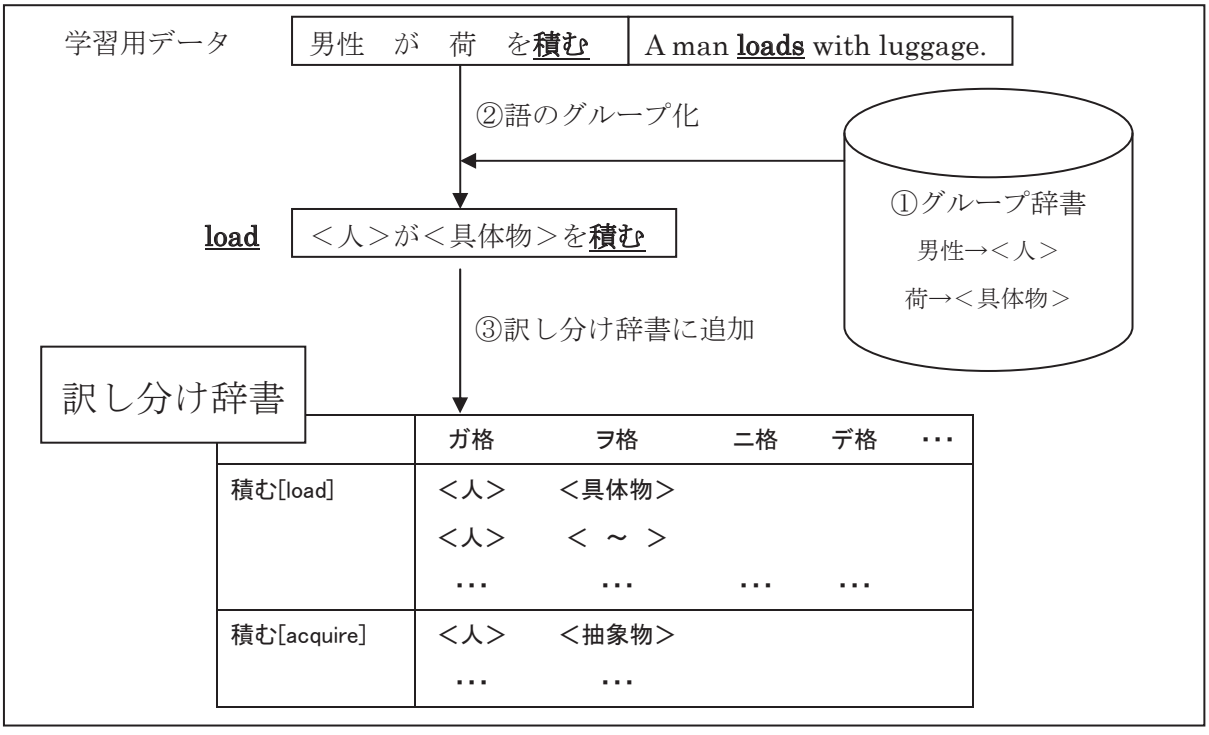

図 1　訳し分け辞書作成フローチャート

表１　対訳が include となる文の翻訳結果

	Infoseek	Google	excite	Yahoo!
include	30	16	12	30
contain	0	10	18	0
その他	0	4	0	0

表２　対訳が contain となる文の翻訳結果

	Infoseek	Google	excite	Yahoo!
include	30	14	19	29
contain	0	15	9	0
その他	0	1	2	1

表１と表２から、翻訳において意味情報を用いている翻訳サイトとそうでないものとがはっきり分かれた。「Google 翻訳」と「excite 翻訳」は訳し分けをしているようであるが、あまり精度は高くないという結果となった。専門用語や「手段１１」「処理回路３２」といった参照番号による解析ミスが目立った。

３.３. 調査２

次に、グループ化を用いた語の訳し分けについての調査を行った。先の調査で扱った文を対象として、日本語語彙大系 [17] に基づいて人手で語の置換を行った。語彙大系の上位から数えて６層目の概念に意味を考慮した上で置換する。複合名詞など名詞が連続して出現した場合、最後の名詞の語彙体系のみを採用する。

置換した文を同翻訳サイトで翻訳し、訳し分けがなされるか調べた。以下、表３～表６に結果の一部を示す。

表３　語のグループ化により改善された例１

置換前	**＜始発無線周波数励起パルス＞**をそれぞれ含んだ	Each of the first train that **includes** a radio frequency excitation pulse
置換後	**＜物理現象＞**をそれぞれ含んだ	Each **contains** the physics

表４　語のグループ化により改善された例２

置換前	**＜予約情報＞**に含まれる検査内容	contents **included** in the Best Inspection
置換後	**＜知識＞**に含まれる検査内容	contents **contained** in the knowledge examination

表５　正しい翻訳が誤った例

置換前	**＜生体活性なバイオセラミックス粉体＞**を含んだ**＜生体内吸収性の多孔質体＞**と、	And in vivo absorption of porous powder **containing** a bioactive bioceramics,
置換後	**＜構造＞**を含んだ**＜構造＞**と、	Structure **including** the structure,

表６　語彙体系を用いた置換の翻訳結果

	google		excite	
	contain	include	contain	include
×→○	10	9	1	3
○→×	6	4	2	0
×→×	7	6	20	15
○→○	7	11	7	12

表３では「始発無線周波数励起パルス」を「物理現象」と置換した結果、「含む」が [contain] と正しく訳し分けされた。専門用語を置換するために、グループ化を自動化する際に未知語推定をする必要がある。表４では「予約情報」を「知識」と置換した結果、「含む」が [contain] と正しく訳し分けされた。「予約」を削り、「情報」を日本語語彙大系の「id=1008 subject= 知識・知恵」を利用して置換した。日本語語彙大系を用いる場合はグループ化をする階層を考慮する必要がある。表５では「生体活性なバイオセラミックス粉体」と「生体内吸収性の多孔質体」を「構造」と置換した結果、「含む」が [contain] と正しい翻訳がされていたものが [include] となり翻訳に誤りが生じた。

表6に全体をまとめた結果を示す。○が正しいもの、×が誤ったものである。Google ではある程度改善されたが、Excite ではほとんど変化がなかった。

4 おわりに

今回の調査で語のグループ化が訳し分け精度の向上に役立つことが分かった。現在、日本語語彙大系を用いた語の自動グループ化に取り組んでいる。まずこのグループ化の精度を調べる予定である。未知語、専門用語が出現した際、グループの推定をする方法は、上記のように、複合名詞の場合には、最後の語をグループ化する方法で対処する予定であるが、その語も未知語である場合の対処法は現在模索中である。また、熟語単位での訳し分けに対応できないため、熟語に対応した訳し分けを行うには辞書の作成方法を改善する必要がある。

参考文献

[1] 横山晶一、高野雄一：特許文の英語への訳し分けと述語の関係、JapioYearbook2010 (2010) pp.274-279

[2] 横山晶一：特許文の英語への訳し分けと格フレームとの関係、JapioYearbook2009 (2009) pp.262-265

[3] 横山晶一：動的シソーラスを用いた特許文の解析システム、科学技術研究費成果報告書 (2007～2009)

[4] 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 (1999) pp.442-448

[5] Shoichi Yokoyama, Yuichi Takano: Investigation for Translation Disambiguation of Verbs in Patent Sentences using Word Grouping, Fourth Workshop on Patent Translation (2011)

[6] 鈴木勘平、横山晶一：動詞の格情報を用いた特許文の解析、情報処理学会第 72 回年次大会論文集 4W-2 (2010)

[7] 網川隆司、梶博行：構文的共起と意味クラスを用いた訳語選択、平成 22 年度 AAMT/Japio 特許文研究会　報告書 (2011) pp.25-30

[8] Ide, Nancy and Jean Veronis: Introduction to the special issue on word sense disambiguation Using Bilingual Comparable Corpora. In Proc. of the 19th International Conference on Computational Linguistics (1998) pp. 411-417.

[9] Dagan, Ido and Alon Itai: Word Sense Disambiguation Using a Second Language Monolingual Corpus. Computational Linguistics, 20(4) (1994) pp. 563-594.

[10] Li, Cong and Hang Li: Word translation disambiguation using bilingual bootstrapping. In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (.2002) pp. 343-351

[11] Vickrey, David, Luke Biewald, Marc Teyssier and Daphne Koller: Word-sense disambiguation for machine translation. In Proc. of the Conference on HLT/HMNLP (2005) pp 771-778

[12]（財）日本特許情報機構：AAMT / Japio 特許翻訳研究会特許情報データベース (2004)

[13] Infoseek マルチ翻訳：(http://translation.infoseek.co.jp/)

[14] Google 翻訳：(http://www.google.co.jp/language_tools?hl=ja)

[15] excite 翻　訳：(http://www.excite.co.jp/world/)

[16] Yahoo! 翻訳：(http://honyaku.yahoo.co.jp/)

[17] 池原悟他編：日本語語彙大系、岩波書店 (1999)