特許文の英語への訳し分けと述語の関係

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

山形大学大学院情報科学専攻　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

特許文の詳細説明や要件の文が長大で難解であり、複雑な構造を持つことはよく知られている。すでに本誌でもこれについては言及した [1]（それ以前の文献についてはこの文献および [2] を参照）。

[1] では、格フレームという、動詞に対してどのような要素が共起するかについての情報を用いて、特許文を英訳した時に、その情報が特に訳し分けに役に立つかを調査した。結果的には、格フレームに書かれている情報の偏りによって、必ずしも訳し分けに寄与するとは言えないという結論が導かれた。

今回は格フレームよりもやや一般的な、述語項構造を用いた場合に、それが訳し分けにどのように寄与するかについての予備的な調査を述べる。

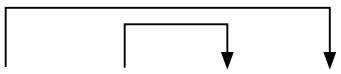
2 述語項構造と格フレーム、結合価の関係

2.1 述語項構造 [3]

述語項構造とは、述語と項（日本語では述語と格関係にある単語）との関係のことで、情報抽出、自動要約、機械翻訳など広範囲のテキスト処理のタスクにおいて、意味処理による精度向上を考えたときにキーとなると考えられている要素技術である。格フレームは、動詞と他の要素との格関係に限られているが、述語項構造は、動詞、形容詞、「A は B だ」といった文まで幅広く扱えることが特徴である。

動詞や形容詞などの「述語」は文の中心的な要素であり、動きや状態などの事態を表す。「名詞＋格助詞」という形式で述語が表す事態に関係する人やものを表現する要素を「項」と呼ぶ。述語項構造は、文中の各述語について述語が表す意味を補う働きをする項を同定する。

例えば、「花子はお弁当を買って食べた。」という文の場合は次のようになる。



例：　花子は　お弁当を　買って　食べた。

買って（ガ格：花子　ヲ格：お弁当）

食べた（ガ格：花子　ヲ格：お弁当）

この文では、述語は「買って」と「食べた」の二つが

あり、どちらもガ格は「花子」、ヲ格は「お弁当」を取る。

文の述語項構造を解析することには、以下のような利点がある。

(1) 情報検索の結果向上

現在の情報検索は、名詞・動詞などの単語の共起だけで検索しているため、入力した人が意図している文が検索されない場合がある。例えば「鍋を粘土で作りたい」というとき、従来の検索では「鍋 粘土 作る」といったような単語の共起で検索する。しかしこの場合「鍋を粘土で作る」以外にも「鍋で粘土を作る」や「鍋と粘土で(何かを)作る」なども検索される。

実際に Yahoo! JAPAN[4] の検索で (a)「鍋を粘土で作る」、(b)「鍋 粘土 作る」で検索したところ、(a) の検索結果では 1 ～ 3 番目は関連したサイトだったが、4 番目は「粘土を鍋で作る」ことが書かれているサイトであった。また、(b) の検索結果では、意図していた「鍋を粘土で作る」ことが書かれていたサイトは 2 番目の結果で、1 番目の結果は (a) の 4 番目に表示されたサイトであった。なお、「粘土で鍋を作る」も検索したが、こちらは (a) の結果とは異なり、1 番目は関連したサイトで、2 番目は「鶏鍋を作る」と「通気孔を粘土で塞ぐ」ことが書かれたサイトであった。

以上から、情報検索に述語項構造解析を組み込み、入力を単語ではなく文に変更して述語項構造を解析することによって、文を意味単位で検索し、より良い情報検索が行える。

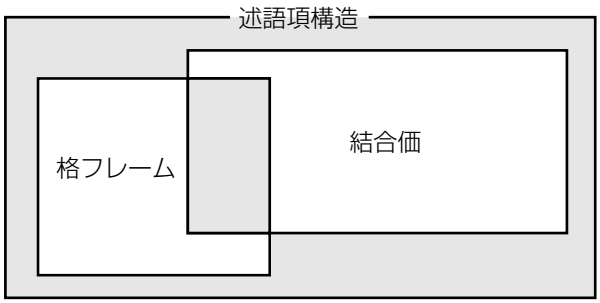


図 1　述語構造と各フレーム、結合価の関係

(2) 文の意味理解への補助

述語項構造は文の意味の骨格を表すものであり、文の述語項構造を表示することによって、その文の意味理解を助けることができる。

述語項構造は、文中の述語に焦点を置き、文の意味の内部構造を表している。そのため、述語が含まれていない、または省略されている文については、内部構造を表せないという問題がある。以下に、例を示す。

例文 A：明日で 20 歳になる。

例文 B：明日で 20 歳。

上記の例文 A と例文 B は、同じ意味を持っている。例文 A は動詞 (述語)「なる」が含まれているため述語項構造で表せるが、例文 B は述語が含まれていないため、述語項構造で表すことができない。

述語項構造は以下で説明する格フレーム、結合価と図 1 で表されるような関係になっている。

2.2　格フレーム [5]

格フレームとは、動詞を基準として語と語の間の意味的関係を記述したものである。格フレームを用いることで次のような構文的曖昧性を解決できると考えられる。

・望遠鏡で泳ぐ少女を見た。

・クロールで泳ぐ少女を見た。

上記の例では動詞「{ 望遠鏡 } で見る」「{ クロール } で泳ぐ」という知識が格フレームにあれば構造を正しく解釈できる。Web 上のテキストから作成されたコーパスを基に格フレームを構築することで、新聞コーパスから構築されたものよりもカバレージが高く、常識的な知識を多く含む利点がある。その反面、一つの用言に対する格フレーム数が多くなりすぎていたり、Web 特有の俗語が見られたりするという欠点がある。

2．3．結合価 [6]

結合価とは、用言を中心として、用言に係る体言と、体言の後に続く格助詞の組み合わせで用言の特徴を記述

したものである。特に、動詞を中心にした結合価とは、「動詞を修飾する単語」「動詞を修飾するときの格関係を表示する格助詞あるいは連語などによる格助詞表現」「動詞」の３つをセットにして、文の形を特徴付ける形式のことである。各用言はそれぞれ必ず取る項の数が決まっている。その例を表１に示す。

表１　結合価における各用言の取る項

取る項の数	動詞	取る項
1	寝る	寝る人
2	蹴る	蹴る人、蹴られるもの
3	貸す	貸す人、貸してもらう人、貸すもの

表１の取る項では「人」や「もの」に限定して表現したが、実際の結合価では｛動物、車｝などのようにその動詞に関係する単語群や、〈歩行する動物〉などのように単語群を意味的なグループ名で表したものとして表現する。また、例えば表１に示した動詞「貸す」の場合は、「貸す人」「貸してもらう人」「貸すもの」の３つの項が必要となる。しかし、日本語の場合はこの必須項が自明である場合は省略する場合がある。例えば、ただ「寝る」と言った場合には、「寝る人」の項には「寝ると言った人」が入る。また、項の数はあくまでも必須の項の数であり、例えば「私は彼女に本を無料で貸す」などといったように、取る項の数を増やすことも可能である。

3 関連研究

3.1　格フレームによる特許文の訳し分け

昨年度の本誌 [1] や、AAMT/Japio 特許翻訳研究会報告書 [7] にも書いてあるが、単語の意味を特定する手が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 [8,9]。

通常、ある動詞に対する格フレームは複数個ある。格フレームの違いは動詞の意味の違いを表現していると考え、格フレームの違いが翻訳された英訳の違いと対応しているならば格フレームを利用した動詞の訳し分けが可能となるとして研究をした。格フレームは Web から自動構築された既存の格フレームである。調査の結果、この格フレームを特許文の訳し分けに利用することは困難であることがわかった。しかし、日本語文を英訳別に分類して格構造の傾向を見ると、ある程度の規則性があり、格フレームのような構造が構築できる可能性がある動詞が存在した。この結果を既存の格フレームに反映させることで現状の問題点を改善し、訳し分けに利用可能な格フレームに修正できるのではないかとの結論を出している。

3.2　述語項構造解析システム

昨年度大澤は、文の意味はその文に含まれる動詞を中心としたものであると考え、格フレームと結合価を用いて、動詞を中心とした述語項構造解析システムを構築した [10, 11]。このシステムは、KNP[12] の係り受け解析、格解析の結果を元に、結合価を利用した学習データを用いて項の同定を行い、述語項構造を表すシステムである。

先行研究である SynCha[13] では、係り受け解析とタグ付けによって項の同定を行っているが、こちらのシステムでは係り受け解析と格解析によって項の同定を行っている。

評価としては、動詞が正しく抽出できた場合は、高い精度をあげることに成功しているが、動詞の抽出がうまく行えなかった文が多く、全体的に精度が高いとは言えない結果となった。課題の解決や動詞の抽出方法を変えることで、より制度の高い述語項構造解析システムを作成できるのではないかと結論づけている。

4 研究内容

4.1 述語項構造データベース

訳し分けの為に述語項構造データベースを作成する。図２にデータベース作成の処理手順を示す。訳し分けの対象となる動詞を含んでいる日本語の特許文を、先行研究の述語項構造解析システムに入力して格の構成について解析する（表２）。そうして出た結果をデータベースに格納し（表３）、訳し分けに用いる。

今回の研究で使用した特許文データは 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文を扱った [14]。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。

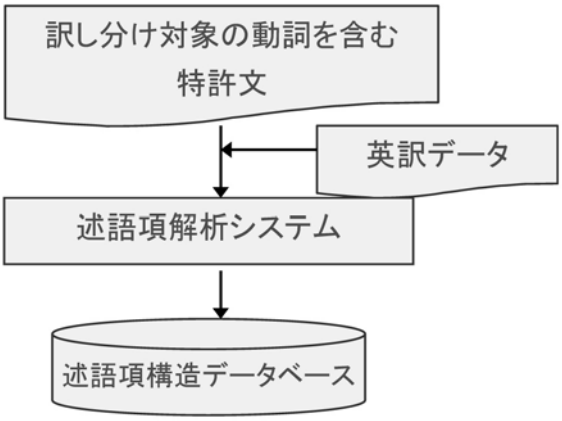


図２　述語項構造データベースの作成

表２　述語項構造の出力例

動詞	ガ格	ヲ格	ニ格	カラ格	ヘ格	ト格	ヨリ格	マデ格	デ格
含む	核酸	配列							

表３　「含む」の述語項構造データベース

含む	include	contain
ガ格	0	6
ヲ格	10	5
ガ格ヲ格	40	10
ガ格ニ格	0	29
ガ格ヲ格ニ格	0	0

4.2 データベースを用いた訳し分け

特許文の翻訳の際には述語項構造データベースを用いて、適切な動詞の訳し分けを判断する。図３に翻訳の際の処理の流れを示す。特許文の述語項構造を解析した後、訳し分け対象の動詞に対してデータベースを参照し、一致する述語項構造の中で出現頻度の高い対訳動詞を正答とする。翻訳システムから出力された文の訳し分けが正しいか判定する。現在の研究では翻訳システムをどのように利用するか検討しているため、動詞の訳し分けを判定するに留まっている。

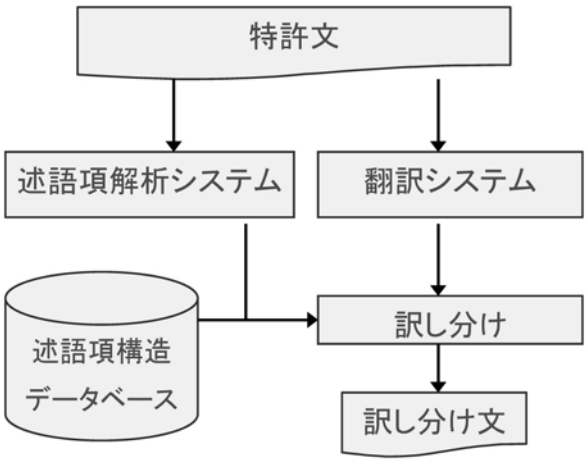


図３　述語項構造を用いた訳し分け

4.3 実験

今回は出現数の多く、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〕）と「得る」（〔acquire: 不断の努力で得る・習得する〕，〔obtain: 努力・要求・懇願で獲

得する〕) を扱った。扱う動詞を限定しているため、先行研究の述語項構造解析システムは 100 文中 81 文が正答と、高い精度で解析が出来ていた。

システム構築の前に手作業で、対象の動詞を含んでいる 50 文の述語項解析を行った。「含む」は表3、「得る」は表4のような結果が得られた。「含む」は訳語ごとに述語項構造に明確な差異があるのに対して、「得る」はどちらの訳も同じような述語項構造となっており、明確な違いは出なかった。検証数が 50 と少ないため、十分な量のデータを収集し、格の構成による訳し分けの有用性についての検証を行う必要がある。

表4　「得る」の述語項構造データベース

得る	acquire	obtain
ガ格	4	6
ヲ格	20	24
カラ格	1	4
デ格	0	8
ガ格ヲ格	7	3
ガ格カラ格	6	1
ガ格デ格	8	1
ヲ格ニ格	4	1
ヲ格デ格	0	1
ガ格ヲ格デ格	0	1

5 おわりに

本稿では、特許文における述語項構造を利用した訳し分けについて調査した。調査の結果、述語項構造を訳し分けに利用出来る可能性があることがわかった。今後は、作成したシステムから多くの動詞についてこの手法が訳し分けに利用出来るか調べる。係り受け関係についても調査を進めて、特許文特有の傾向がないか調査する予定である。

[参考文献]

[1] 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

[2] 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)

[3] 飯田龍、小町守、乾健太郎、松本裕治：日本語書き言葉を対象とした述語項構造と共参照関係のアノテーション　NAIST テキストコーパス開発の経験から、言語処理学会第 13 回年次大会発表論文集 (2007) pp.282-285

[4]Yahoo!JAPAN：(http://www.yahoo.co.jp/)

[5] 河原大輔、黒橋禎夫：格フレーム辞書の漸次的自動構築、自然言語処理 Vol. 13, No.2 (2005) pp.109-131

[6] 荻野孝野、小林正博、井佐原均：日本語動詞の結合価、三省堂 (2003)

[7] 横山晶一：特許文の訳し分けと動詞の格情報との対応に関する調査、平成 21 年度 AAMT/Japio 特許翻訳研究会報告書 (2010) pp.61-66

[8] 鈴木勘平：動詞の格情報を用いた特許文の解析、山形大学大学院理工学研究科修士学位論文 (2010)

[9] 鈴木勘平、横山晶一：特許文の訳し分けにおける格フレームの有効性、情報処理学会第 72 回全国大会 (2010) 4W-2

[10] 大澤有美：結合価と格フレームを取り入れた述語項構造解析システム、山形大学工学部卒業論文 (2010)

[11] 大澤有美、横山晶一：結合価と格フレームを取り入れた述語項構造解析システム , 平成 21 年度 第

6 回情報処理学会東北支部研究会 (2010) A-1-3

[12] KNP : http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

[13] SynCha : http://cl.naist.jp/~ryu-i/syncha/

[14] (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)