特許分類を利用した分野別対訳知識の抽出

株式会社東芝　研究開発センター
熊野　明

PROFILE 1982年東京工業大学工学部卒業。同年東京芝浦電気(株)(現、(株)東芝)入社。以来、機械翻訳、電子化辞書などの自然言語処理技術の研究開発に従事。現在、研究開発センター知識メディアラボラトリー主任研究員。情報処理学会、人工知能学会、言語処理学会会員。

akira.kumano@toshiba.co.jp　044-549-2239

1 はじめに

日本の特許情報を海外から検索するニーズは、特許の国際出願の増加に応じて年々高まっている。しかし、日本語の技術文書が理解できる欧米人口は非常に少なく、日本語の特許文書を英語などに翻訳することが必須である。特許庁では、抄録データを人手で翻訳し、PAJ（Patent Abstract of Japan）として公開している。しかしその抄録は特許文書の一部に過ぎず、その他の明細書のほとんどの部分は、機械翻訳システムを利用して日英翻訳した結果を提供している[1]。

特許文書は、新しい技術を含むという性質上、新語や専門用語が多く含んでいる。これらの用語に対して事前に十分な機械翻訳辞書を準備することは不可能である。この問題を解決するため、特許抄録とPAJのような対訳文書データを利用して、機械翻訳に利用可能な対訳知識を抽出する技術をすでに開発した[2][3]。この対訳知識抽出技術は、形態素解析および構文解析で明確に判定できる用語のうち、(1)全くの新語（すなわち、未知語）、(2)専門用語と思われる語（名詞連続などの複合語）を抽出し、その後、対応訳文の情報を利用して抽出用語の訳語推定を行うもので、The翻訳シリーズの辞書構築支援技術としてすでに製品化している[4]。また、本技術の製品化後も、特許文書の特徴を利用して精度の改良を続けている[5]。

ところが、特許文書の翻訳精度を向上させるためには、新語や専門用語以外に頻出する一般用語の訳語精度も重要である。しかし、一般用語は専門分野を問わず幅広く使われるため、分野によって異なった訳語が必要な例も少なくない。

そこで本研究では、新語や専門用語以外の一般用語に関して対訳知識を抽出する方法を実証する。さらに、分野別の対訳知識抽出の有効性を示すため、特許抄録に付与されている特許分類コード（IPC）を利用して対訳文書を分類し、分野別に対訳知識の抽出を行い、その有効性を確認する。

2 特許抄録とPAJに含まれる対訳知識

図1は特許抄録とそのPAJの例である。

特願 2002-6933
【発明の名称】
小型電子計算機
【課題】
装置本体の薄型化を損なわずに、効率良く冷却することができる小型電子計算機を提供する。
【解決手段】
ノート型パーソナルコンピュータであって、**装置**本体１と、この**装置**本体１にヒンジを介して開閉可能なディスプレイ部２と、**装置**本体１の上面に設けられたキーボード３などから構成され、ディスプレイ部２を開くことにより、**装置**本体１内に収納していたファン（軸流ファン）７を自動的に立ち上げるとともに、上面カバー８及び背面カバー９も自動的に上方に上げて**装置**本体１内に空間を作り、ファン７によって**装置**内部の熱せられた空気を吸い込み、**装置**外部へ熱を効率良く排出することができる。また、ディスプレイ部２を閉めると逆の動きをすることにより、ファン７は自動的に倒れて**装置**本体１内に収納でき、上面カバー８及び背面カバー９も自動的に閉じることが可能となる。

[PAJ]
TITLE:
SMALL SIZED COMPUTER
PROBLEM TO BE SOLVED:
To provide a small sized computer that can cool itself efficiently maintaining the thinness of the main body of a **device**.
SOLUTION:
This is a notebook personal computer that comprises a **device** body 1, a display section 2 that is connected to the **device** body 1 with a hinge and able to open and close, and a keyboard 3 set at the top of the **device** body 1, and opening of the display section 2 starts up a fan 7 (an axial flow fan) automatically, and also raises a top cover 8 and a back cover 9 upward and makes a space inside the **device** body 1, and the heated air inside of the **device** can be inhaled and exhausted to outside of the **device** efficiently by the fan 7. Closing of the display section 2 makes the movements in a opposite way and brings down the fan 7 automatically, houses it in the **device** body 1 and closes the top cover 8 and the back cover 9.

図1　日本語特許抄録とその英訳の例（その1）

現行の日英機械翻訳エンジンにおいて「装置」の第1訳語はequipmentであるが、上の特許文書ではdeviceを訳出すべきであることがわかる。つまり、「装置」に対する訳語deviceの情報を獲得する必要がある。このように、新語や専門用語だけでなく、一般的に使われる語の訳語精度が特許などの専門文書の翻訳に重要であることを示している。

図2は別の特許抄録とそのPAJの例である。

特願 2002-253207
【発明の名称】
組織凍結保存法
【課題】
細胞を、消化酵素などによる細胞分散処理や細胞培養を行わずに、細切**組織**のまま、血清を含まないガラス化液で凍結保存する方法を提供する。
【解決手段】
生体からバイオプシー針で採取した**組織**、または死後間もない死体や屠畜場などで採取した**組織**を鋏でペースト状になるまで細切し、燐酸緩衝液で良く洗浄する。細切**組織**を血清を含まないガラス化液に入れ、細切**組織**浮遊液とし、これを個体識別ラベルを貼った凍結保存用容器に入れ、密封する。これを液体窒素上で冷却後、液体窒素中に入れ、液体窒素中で保存する。

[PAJ]
TITLE:
TISSUE FREEZE-PRESERVATION METHOD
PROBLEM TO BE SOLVED:
To provide a method for freeze-preserving cells directly in the form of minced **tissue** in serum-free vitrified liquid without a cell dispersion treatment with digestive enzymes and/or cell culture.
SOLUTION:
This method for freeze-preserving cells comprises the following procedure: a **tissue** collected with a biopsy needle from a living body or collected from corpse soon after death at a slaughterhouse or the like is minced with scissors into a pasty form, which is then thoroughly washed with a phosphate buffer solution; the resultant minced **tissue** is put into a serum-free vitrified liquid into a minced **tissue** suspension, which is then put into a freeze-preserving container stuck with an individual body identification label and sealed up; the resultant container is chilled on liquid nitrogen, put thereinto, and then preserved in the liquid nitrogen.

図2　日本語特許抄録とその英訳の例（その2）

現行の日英機械翻訳エンジンにおいて「組織」の第1訳語はorganizationであるが、上の特許文書ではtissueを訳出すべきであることがわかる。しかし、「組織」をtissueと訳出するのは、特許文書一般的な現象ではなく、おそらくこの特許を含む分野特有の現象であると思われる。

3 実験とその結果

未知語・複合語を対象として製品化した用語抽出・訳語推定エンジンを一部改良し、一定の一般用語に対しても訳語推定できるようにした。今回対象とした一般用語は、次の10語である。分野を問わず特許文書に多用される用語である。

情報、方法、手段、装置、構成、制御、構造、機構、合成、組織

今回さらに、分野による訳語知識の違いを調べるため、特許に付与されているIPC（国際特許分類）コードをもとに、分野ごとに特許対訳文書を集めて処理することにした。

2002年の国内出願特許374,700件には、それぞれ複数のIPCコードが付与されている。各特許の先頭の

IPCコードによって分類した分布を表1に示す。

今回の実験では、上位の8分野のほか、異なった分野による違いを調べるために、C08L, F16H, E04B, C12Nの4分野を加えた12分野を選択した。

各分類から約100件の特許抄録対訳データを取り出し、一般用語の対訳知識を抽出できる方法で、用語抽出・訳語推定を行った。実験では、固定の訳語が決まっているものを処理する意味はないので、日本語抄録中にある「発明の名称」、「課題」、「解決手段」、PAJ中の「TITLE:」、「PROBLEM TO BE SOLVED:」、「SOLUTION:」の固定見出しは除いて行った。

訳語推定結果を表2に示す。

推定訳語は、第3候補まで出力したものである。第1推定訳語に対しては、正しいと判断したものに○、それ以外のものに△を付けた。最下欄の正解率は、第1推定訳語が正しい割合を分野ごとに示したものである。

表1　2002年国内出願特許のIPC分類別件数（一部）

件数順	先頭IPC	出願件数
1	**G06F**	30,057
2	**H01L**	16,631
3	**H04N**	12,358
4	**G11B**	8,549
5	**G03G**	7,700
6	**G02B**	6,863
7	**A63F**	6,822
8	**B41J**	5,670
9	G01N	5,186
10	H04L	4,949
11	B65D	4,831
12	H01M	4,825
13	H05K	4,317
14	B29C	4,230
15	H04M	3,823
16	**C08L**	3,659
…	…	…
32	**F16H**	2,210
33	**E04B**	2,201
…	…	…
62	**C12N**	1,471
…	…	…

4 結論と今後の課題

特許対訳文書から、一般用語の訳語情報を自動抽出することができた。

また、「装置」の訳語情報が分野によって、G06F、H04Nなどでは"device"、B41Jなどでは"apparatus"、H01Lでは"system"と異なり、一部の分野（C12N）だけに「組織」に対する"tissue"の訳語情報が抽出されるなど、分野による訳語の特徴も抽出することができた。12分野の抄録対訳テキスト各100対から抽出した実験では、10種類の用語全体に対して、抽出できたもの81種類のうち、正しい訳語知識であるものは72件で、正解率は約82%であった。未知語・複合語を抽出対象にした報告[5]では、同程度の特許対訳抄録を使って未知語14%、複合語69%の精度を確認しているので、相対的に高い精度である。

また、この結果はそのまま機械翻訳知識にするものではなく、人手のチェックを行ってから辞書化するものであるので、実用レベルの精度であると考える。

しかし、一般用語の訳語情報という意味では、現段階でも改良が可能な部分もある。

形態素解析・構文解析結果から用語を抽出しているので、複合語として辞書にあるものは、単語としての一般用語とその訳語情報が抽出できない。たとえば、G06F分野における単語「制御」の頻度は3であるが、実際には他に次のような用語が出現している。

ディジタル**制御**装置
告知**制御**手段
通信**制御**回路
電源**制御**システム
カード**制御**部

これらは、「ディジタル」＋「制御装置」、「告知」＋「制御手段」、「通信制御」＋「回路」、「電源制御」＋「システム」、「カード」＋「制御部」のような語構成をしており、形態素解析結果から「制御」単独の単語を抽出し

表2　特許抄録に含まれる一般用語の分野別訳語推定結果

用語	A63F分野		B41J分野		C08L分野	
	頻度	推定訳語	頻度	推定訳語	頻度	推定訳語
情報	22	○ information	10	○ information	0	−
方法	11	○ method − how	38	○ method	10	○ method − process − using
手段	40	○ means − measuring	43	○ means	0	−
装置	35	○ device − apparatus − slot device	57	○ apparatus − device − imaging apparatus	9	○ device − apparatus − forming device
構成	5	○ structure	8	○ constitution − constituted − structure	1	△ formed
制御	3	○ control	27	○ controlling	0	−
構造	2	○ structure	4	○ structure	3	○ structure
機構	14	○ controls	3	○ mechanism	0	−
合成	0	−	1	○ synthesizing	0	−
組織	0	−	0	−	0	−
正解率		8 / 8		9 / 9		3 / 4

用語	C12N分野		E04B分野		F16H分野	
	頻度	推定訳語	頻度	推定訳語	頻度	推定訳語
情報	1	○ information	0	−	0	−
方法	66	○ method − using	12	○ method	0	−
手段	3	○ means	1	○ means	0	−
装置	5	○ apparatus − instrument − equipped	5	○ device	12	○ device − apparatus − unit
構成	1	△ comprises	0	−	0	−
制御	1	△ biosynthesis − phenotype − transduced	0	−	0	−
構造	2	○ structure	9	○ structure	8	○ structure
機構	1	○ mechanism	0	−	8	○ mechanism
合成	3	△ semisynthetic − medium − agar	0	−	0	−
組織	7	○ tissue − forming	0	−	0	−
正解率		7 / 10		4 / 4		3 / 3

表2　特許抄録に含まれる一般用語の分野別訳語推定結果（つづき）

用語	G02B分野		G03G分野		G06F分野	
	頻度	推定訳語	頻度	推定訳語	頻度	推定訳語
情報	8	○ information	0	−	140	○ information
方法	14	○ method- using − aligning method	31	○ method- using − process	80	○ method − using
手段	11	○ means	6	○ means	70	○ means − method − processing method
装置	19	○ device − instrument − lens device	41	○ device − system − forming device	67	○ device − system − processing means
構成	4	△ constituted − structured − comprising	3	○ formation − structure − element	5	△ constituted − form
制御	0	−	4	○ controlling	3	○ control
構造	7	○ structure	0	−	2	○ structure
機構	0	−	2	○ mechanism − structure	1	○ mechanism
合成	0	−	0	−	1	△ information − personal − pattern
組織	0	−	0	−	0	−
正解率		5 / 6		6 / 6		7 / 9

用語	G11B分野		H01L分野		H04N分野	
	頻度	推定訳語	頻度	推定訳語	頻度	推定訳語
情報	48	○ information	2	○ information	34	○ information
方法	53	○ method	24	○ method	39	○ method − process − control method
手段	12	○ means	15	○ means − stored means	23	○ means − way
装置	36	○ device	12	○ system − apparatus	70	○ device − units − system
構成	4	△ formed - constitution	1	○ configuration − semiconductor − manufacturing	5	○ component − structure
制御	3	○ control	3	○ controlling	10	○ control
構造	7	○ structure	3	△ structured	0	−
機構	5	○ mechanism	0	−	0	−
合成	2	○ synthesis - composed	0	−	0	−
組織	0	−	0	−	0	−
正解率		8 / 9		6 / 7		6 / 6

ていない。「制御」を含む（複合語としての）用語、たとえば「ディジタル制御装置」に対しては、用語抽出・訳語推定を行うが、より一般的に「制御」の訳語推定を行う際には利用していない。

今後は、これらの課題を解決することにより、実際に分野別特許文書からの対訳知識を抽出し、機械翻訳精度が向上することを確認する予定である。一定の精度向上が確認できた後、今回実験した手法を機械翻訳用対訳知識の構築支援に活用するとともに、機械翻訳ユーザ向けの辞書構築支援技術として実用化を目指す。

参考文献

[1] 特許電子図書館、工業所有権情報・研修館、http://www.ipdl.ncipi.go.jp/
[2] 熊野、平川: 対訳文書からの機械翻訳専門用語辞書作成、情報処理学会論文誌35巻、11号、pp.2283-2290、(1994)
[3] 熊野: カタカナ表記からの英訳推定による専門用語辞書作成、言語処理学会 第1回年次大会、pp.221-224、(1995)
[4] 英日/日英翻訳ソフト「The翻訳シリーズ」http://hon-yaku.toshiba-sol.co.jp
[5] 熊野: 特許文書の特徴を利用した対訳知識抽出、Japio 2006 Year Book、(2006)