

意味的用法に着目した日本語名詞の英訳語選択について

池原 悟[†] 村上 仁一[†] 桐澤 洋[†]

日英機械翻訳において精度良い名詞の訳語選択技術を実現するため、日本語の基本的な名詞の英訳語の多義構造について調査し、名詞の「意味的用法」に関する分類知識（文中での名詞の用法の意味的な分類）を使用した訳語選択の可能性について検討した。具体的には、まず、IPAL 辞書に収録された名詞 1,082 語を対象に、和英辞書を使用して名詞の持つ英訳語数の静的な分布を調べると共に、新聞記事 1 万文の対訳コーパスを対象に、使用頻度の高い名詞についての訳語数の動的な分布を調べた。次に、複数の英訳語を持つ名詞（「訳語多義語」と言う）を対象に、「意味的な用法」として「日本語語彙大系」で定義された「一般名詞意味属性」（単に「意味属性と言う」）を使用した場合の訳語選択精度を調べた。また、その結果を基に「意味的用法」を使用した名詞訳語選択の限界と、それ以外で必要と考えられる知識について考察した。

その結果によれば、IPAL 辞書の名詞では複数の訳語を持つものは約 5 割（499 語）で、そのうち 55% は「意味属性」の知識によって一意に訳語が決定できること、また、それを含む 87% は、訳語候補が絞り込めることが分かった。訳語多義語の平均多義数は 3.02 から 1.74 に減少し、訳語正解率は 38.7% から 78.6% に向上する。新聞記事で使用頻度の大きい名詞では、訳語多義数の平均が 3.0 から 1.4 に減少し、訳語正解率は 53.8% から 92.2% に向上する。なお、失敗用例の分析では「意味属性体系」の改良で最大約 6% 程度の精度向上が見込まれることから、これが名詞の「意味的用法」の知識を使用した方法の限界と推定される。

キーワード：機械翻訳，訳語選択，意味的包含関係，単語意味属性，意味解析

English Word Selection for Japanese Nouns using Semantic Usage

SATORU IKEHARA,[†] JIN'ICHI MURAKAMI[†]
and HIROSHI KIRISAWA[†]

Translation word selection, especially for nouns, is one of the most important issues in machine translation. This paper studied the capability of English word selection for basic Japanese nouns using information of semantic usage. In this study, basic nouns of 1,082 words registered in IPAL dictionary were used for evaluation. The results are as follows. Translation of 55% out of 499 words which have multiple translation words can uniquely be determined by using the information of semantic usage. The number of translation candidates can be reduced for 87% of 499 words. In the study of newspaper article translation, the average value of the word ambiguity can be reduced from 3.0 to 1.4 by this method.

1. はじめに

機械翻訳の訳文品質の向上を図る上で名詞や動詞の訳語選択は、重要であると同時に大変難しい問題の一つである¹⁾。特に日本語と英語のように言語族の異なる言語間では、対象の捉え方の違いが、この問題をより困難にしていると考えられる。ちなみに日英機械翻訳における訳文品質の分析²⁾では、訳語選択の不適切さが、訳文品質低下の原因の 40% を占め構文解析の誤り (25%) を抜いて最大であると報告されている。

訳語選択の精度を向上させるには、文中で使用され

た単語の意味（語義）を解析することが必要である。単語の意味については、共起する単語の組で定義する試み³⁾もあるが、単語の語義が、通常、辞書で定義されることを考えると、文中の各単語が辞書上で定義された語義のうち、どの語義で使用されているかを判定することが必要である。この問題に対し、文中での他の単語との共起関係から判定する方法として、統計的な方法⁴⁾ やシソーラスを使用した方法⁵⁾⁶⁾ が着目され、ベクトル空間法の応用⁷⁾⁸⁾ が試みられている。他にも、「重要な内容は、談話文中、言い方を変えて何回も表現される」点に着目し、文の冗長性を手がかりに意味解析を行う方法⁹⁾ や、さまざまな知識を組み合わせる方法¹⁰⁾ も提案されている。

また、動詞に的を絞った意味解析の方法としては、

[†] 〒 680-8552 鳥取市湖山町南 4-101
鳥取大学工学部 Tottori University
E-mail: ikehara@ike.tottori-u.ac.jp

共起レベルと頻度をパラメータとする尤度と辞書的記述を基にした制約を用いる方法¹¹⁾ や単語の直接的な共起関係を用いる方法¹²⁾ がある。最近では、ベクトル空間法を応用する方法¹³⁾¹⁴⁾ や格文法を応用する方法^{15)~17)} の研究が盛んであり、意味解析のための辞書記述法の研究¹⁸⁾ も行われている。意味解析に使用する格パターンを自動学習する方法としても、いくつかの研究^{19)~22)} が行われてきた。

これらの研究のうち、統計的学習による方法は、学習用として大量の標本データを必要とすることが問題となっている。格文法を使用する方法でも文法規則の整備が容易でないことが問題であったが、最近、日英言語対を対象に結合価パターン対辞書(「日本語語彙大系」)が開発された²³⁾。日英機械翻訳の場合、動詞の意味を正しく翻訳するには、約 2.5 万件のパターンが必要²⁴⁾ と言われているが、この辞書では日本語動詞の訳し方が 1.4 万件の結合価パターンによって定義されており、動詞の訳語選択の品質は大幅に向上した。

これに対して、名詞の意味解析では、談話解析で得られた場面情報によって、英語名詞の多義解消を試みたもの²⁵⁾ や連体修飾句の係る名詞を対象に格の情報と意味属性を用いる方法²⁶⁾ が提案されている。また、最近では、名詞辞書における多義構造の記述に関する研究²⁷⁾²⁸⁾ もあるが、対象は限定的で、有効な方法が見つかっていない。

そこで、本稿では、名詞の日英訳語選択の技術的問題点を明らかにするため、日本語の基本的な名詞を対象に英訳語の多義構造について調査すると共に、名詞の意味的な使われ方(「意味的用法」)に関する知識に着目した英訳語選択の可能性について検討する。具体的には、まず第 1 に、IPAL の基本名詞²⁹⁾ を対象に、英訳語数の静的な分布を調べると共に、新聞記事 1 万文を対象に英訳語数の動的な分布を調べる。第 2 に、それらの名詞の中で複数の英訳語を持つ名詞を対象に、「日本語語彙大系」で定義された名詞の「意味的用法」の知識(「一般名詞意味属性体系」)を使用した場合の訳語選択の効果を調べる。最後に、これらの結果を基に方法の限界と名詞の訳し分けが必要と見られるその他の知識について考察する。

2. 日本語名詞の英訳語多義

2.1 対象とする名詞の種類

(1) 日本語名詞の種類と訳語選択

文単位の日英機械翻訳を考えたとき、日本文中の名詞は、必ずしも同じ名詞の英語単語に訳されるとは限らず、異なる品詞の単語として訳される場合や対応す

る英単語の現れないような場合もあり、訳し方はさまざまである。本稿では、主として日本文中の名詞が英文の中で名詞に訳される場合を対象とする。

ところで、日本語名詞を表 1 のように分類すると、すべての名詞を統一的に扱うのは困難であることが分かる。例えば、表 1 で固有名詞は、その数も多いが、人名、地名などの区別ができれば、辞書によって訳語は、ほぼ一意に決定できる。形式名詞は、出現頻度は高いため、その翻訳方式は重要であるが、機能語的な用法が多く、その用法は多彩である。幸い語数は少ないこともあって、既に個別的な検討が行われている²⁹⁾。代名詞の場合は、指示対象が特定されなくても、英訳語は、日本語とほぼ 1 対 1 に対応する。また、時詞は、副詞的な用法が多いため、形容詞、副詞など修飾要素の翻訳方式と合わせて検討することが必要である。これに対して、一般名詞、用言性名詞、転生名詞は、英文中、名詞に訳される場合が多いと考えられる。以上から、本稿ではこれらの名詞を中心に検討する。

(2) 検討対象とする IPAL 基本名詞とその訳語

本稿では、「計算機用日本語基本名詞辞書 IPAL」(以下では、「IPAL 辞書」と称す)³⁰⁾ に収録された名詞(1,082 語:以下では「IPAL 基本名詞」、又は、単に「基本名詞」と称す)を検討対象とする。この辞書には、語数は少ないが、本稿で対象とする一般名詞、用言性名詞、転生名詞の基本的な語が多くを占める。

また、これらの名詞に対する英語訳語としては、日英機械翻訳システム ALT-J/E³¹⁾ の日英対照辞書(以下では、単に「和英辞書」と言う)に収録された訳語を使用する。この辞書に収録された英語訳語は、三省堂、ライトハウスなどの市販の和英辞書から得られた訳語を編集したもので、専門用語としての訳語や慣用的な表現となる場合を除き、通常の文の翻訳で使用される訳語がほぼ収録されている。また、名詞の意味属性と英訳語との対応関係も示されている。

ところで、IPAL 辞書に収録された名詞は、いずれも、見出し語はかな書きされているのに対して、日英対象辞書に収録された語の見出し語は、通常の和英辞書と同様、かな漢字交じりの標準表記となっている。そこで、本稿では、IPAL 辞書の 1,082 語に対応するかな漢字交じり語 1,144 語を検討の対象とする。

IPAL 辞書の名詞は、必ずしも日本語の基本名詞を網羅しているとは言えない。また、若干の代名詞(5 語)、形式名詞(1 語)、固有名詞(1 語)が含まれるが、大多数は一般名詞、用言性名詞、転生名詞の何れかであり、本研究の目的に合致する。

表 1 名詞の種類と訳語選択
Table 1 Noun Types and Their Translations

大分類	小分類	単語の例	訳語選択技術へのアプローチ
一般名詞	普通名詞	山, 人間, リンゴ	語数が多く, 体系的な扱いが必要
	副詞型名詞	新規, 前, 向こう	形容詞, 副詞の翻訳方法としての検討が必要
	連体詞型名詞	大型, 同じ, 新型	
用言性名詞	サ変動詞型	計算, 上昇, 理解	語数が多く, 体系的な扱いが必要
	形容動詞型	静か, 自由, 堂々	語数は比較的少ないが, 体系的扱いが必要
転生名詞	動詞転生型	ずれ, 遊び, 動き	
	形容詞転生型	大きさ, 高さ, 深さ	
時詞	—	昨日, 昨日, 未来	副詞的な翻訳が多く, 別途検討が必要
代名詞	人称代名詞	私, 彼, 彼女	語数は少ないが使用頻度は大きい.
	指示代名詞	あれ, これ, それ	訳語は, ほぼ一意に決定可能
形式名詞	—	の, もの, こと	文法的意味の用法が多彩で, 個別的検討が必要
固有名詞	(人名, 地名, 組織名, その他)		語数は大. 訳語は, ほぼ一意に決定可能

2.2 IPAL 基本名詞の訳語多義の構造

(1) IPAL 基本名詞の語義数と英訳語数の関係

IPAL 辞書では, 日本語意味解析の目的で名詞の意味(「語義」)が細かく分類されているが, 英語訳語の選択を目的とする場合は, 和英辞書に登録された英語訳語に対応づけられる程度の細かさで分類されていなければならない。

そこで, IPAL 基本名詞の語義数と英訳語数の関係を調べるため, IPAL 辞書に収録された日本語の語義数と和英辞書に収録された英訳語数を比較すると, 両者が一致する語が 38%, 日本語の語義数の方が多い語が 39%, 英訳語数の方が多い語が 23%であった。また, 日本語の語義数平均が 2.13(最大語義数は 18)であるのに対して, 和英辞書の異なり訳語数は, 平均 1.88 語(最大 12 語)であり, 和英辞書の訳語数の方が平均して若干少ない。これは, 日英機械翻訳では, 一般的な日本語意味解析よりも, 若干, きめの荒い意味分類で良いことを意味する。

(2) IPAL 基本名詞の英語訳語数の分布

IPAL 基本名詞 1,144 語について和英辞書に収録された英訳語数の分布を表 2 に示す。表 2 より, 検討対象とする 1,144 語のうち, 645 語は単一の訳語しか持たないから, 訳語選択が問題となるのは 2 語以上の英訳語を持つ名詞 499 語であることが分かる。また, 最大訳語数は 12 であり, 2 語以上の訳語を持つ名詞の平均訳語数は, 2.98 である。

ところで, 動詞の翻訳方法の研究³²⁾によれば, 日本語用言の一般文型(慣用表現を除く)として, 和語系動詞の場合, 訳語 10~16(最大 6)が 7 語, 漢語系動詞の場合, 訳語 6~7(最大)が 8 語, また, 形容詞では, 訳語 6~9(最大)が 12 語であることが報告されている。これと表 2 を比べると, 用言の場合と比べて名詞の場合, 訳語数の大きい語の数は若干少ない目である。

3. 「意味的用法」による訳語選択

3.1 「意味的用法」による訳語選択の方法

(1) 訳語選択条件としての「意味的用法」

文中で使用された名詞の訳語を決定するには, その名詞がどのような「語義」で使用されているかを知ることが大切と考えられる。これに対して本稿では, 名詞の「意味的用法」の知識を使用することを提案する。

ここで, 名詞の「語義」と「意味的用法」の関係について考えると, 一般に「語義」は, それぞれの単語が表す対象概念として, 通常, 辞書で定義されているのに対して, 「意味的用法」は, 文中での用法を示もので, 「日本語彙大系」³¹⁾で導入された概念である。

例えば, 名詞「学校」は, 本来の意味(「語義」として「教育の機関」を表すのに対して, 「学校に避難する」の文では, 「場所」を示す言葉として使用されている。また, 「学校を建てる」では, 「学校」は「建物」を示す言葉として使用されている。従って, 「意味的用法」は必ずしも「語義」とは一致しないが, 「意味的用法」が決まると「語義」が決まる可能性が大きいと予想される。

例えば, 日本文「私は, 「鶏」を食べる」と「朝, 「鶏」が鳴く」を比べると, 前者は, <食料>としての鶏であり, 後者は, <動物>としての鶏である。そこで, 「鶏」の「意味的用法」として, <食料>と<動物>の用法を定義したとすると, 文中の前後関係によって, いずれの用法であるかが判定できれば, 「hen」と「chicken」の訳し分けができる。また, 「日本の「田舎」は, 寂れている」, 「私は, 連休で「田舎」に帰った」では, 前者の「田舎」の「意味的用法」が<村落>であるのに対して後者の「田舎」の「意味的用法」は<郷里>であることが分かれば, 「country」と「home」の訳語が区別でき

表 2 IPAL 基本名詞に対する英語訳語数の分布
Table 2 Distribution of English Translation Words for IPAL Nouns

訳語数	1	2	3	4	5	6	7	8	9	10	11	12	合計
語数	645	254	116	66	33	16	4	3	3	1	1	2	1,144 語
多義 なし	訳語多義のある名詞 (2 語以上の訳語を持つ名詞) (語数 = 499 語, 平均訳語数 = 2.98 語)												

る .

以上から、複数の訳語を持つ名詞 (以下では「訳語多義語」と言う) の場合、各名詞の「意味的用法」を分類し、各意味的用法に訳語を対応関係を記述しておけば、訳語が決定できる可能性がある .

(2) 名詞の抽象度と「意味的用法」の関係

単語の「意味的用法」を記述するための必要条件として、「意味的用法」の分類精度 (粒度) の問題について考える .

一般に言語表現では、話者の個性性と普遍性を表現するのに適した抽象度の言葉が使用される . 例えば、「鯛」は、対象を「鯖」や「ヒラメ」と区別するとき使用されるが、これらを区別する必要のない時は、より抽象度の高い「魚」が使われる。「魚」は「虫」や「鳥」と区別するとき使用されるが、これらを区別する必要のない時は、「動物」が使用される . さらに「動物」と「植物」の区別が必要でないときは「生物」が使われる . また、行動を固定してとらえ、客体化した概念を表現する名詞として「登り」、「行き」などの名詞があるが、これを区別しないときは「動き」が使用される .

これと同じように、訳語選択では文中で使用される名詞の意味的な制約条件 (「意味的用法」) を文脈上必要とされる最小限の抽象度で記述することが望まれる .

(3) 「意味的用法」として使用する意味属性体系

上記の訳語選択方式を実現するには、名詞の訳語選択のための名詞の「意味的用法」を体系化することが必要であるが、このような知識ベースを新規に開発することは、コスト、期間の点で容易でない、そこで、本稿では、「日本語語彙大系 (第 1 巻)」²³⁾³²⁾ で定義された「一般名詞意味属性体系」(以下では単に「意味属性」と言う) を代用する . この体系は、名詞との共起関係に着目して用言を訳し分ける目的で開発された

「意味的用法」は、文中での単語の意味的な機能に注目したもので、単語が何を表現するための言葉として使用されているかを示すための言葉として「日本語語彙大系」31) で導入された概念である . 例えば、「学校に進学する」、「学校に避難する」、「学校を建てる」の文の場合、名詞「学校」は、それぞれ「組織」、「場所」、「建物」を表す言葉として使用されているから、それらが「意味的用法」である「語義」と混同されやすいので、注意が必要である .

ものであるが、下記の通り、名詞の意味的な用法が詳細に分類体系化されている .

- (1) 名詞の「意味的用法」が、2,710 種類の意味属性でかなりきめ細かく定義されている .
- (2) 意味属性間の意味的な包含関係 (is-a 関係) が、木構造にまとめられている .

3.2 訳語多義から見た名詞の分類

本節では、IPAL 基本名詞 1,144 語のうち、訳語多義語 499 語を対象に、意味属性による訳語選択の可能性について検討する . 意味属性による訳語選択では、文中で使用された名詞の意味属性に対して、英訳語がどれだけ一意に対応するかが問題となる . そこで、訳語多義語について、その訳語と意味属性との関係を調べたところ、訳語多義語は、表 3 に示すように 4 通りに分類された .

<分類 1> 訳し分け可能

複数ある訳語候補すべてに対して、異なる意味属性が対応する名詞は、英訳語を一意に決定することができるから、これを「訳し分け可能」とする .

例えば、表 3 の名詞「貝」に対する英訳語の意味属性を見ると、「shellfish」の場合は、<魚>と<魚介類>が、「shell」の場合は<殻>が対応し、両者は互いに異なっている . 従って、文中でどちらの意味で使用されたとしても、訳語は、意味属性によって区別できる . 名詞「スキー」の場合も同様である .

<分類 2> 場合により訳し分け可能

文中での名詞の用法により、訳し分けが可能になったり不可能になったりする名詞を「場合により訳し分け可能」とする .

例えば、表 3 の名詞「委員」では、「committee」が、<成員>と<複数>の意味属性を持つのに対して、「member of committee」は、<成員>と<単数>の意味属性を持つ . 従って、文中での名詞「委員」の「意味属性」が、<単数>であれば、「member of committee」が選択され、<複数>であれば、「committee」が選択されるが、<成員>の場合は、両者を区別することができない . また、名詞「脱線」の場合は、意味属性が、<目的>、<話>、<事件>の何れかの場合は、訳語は一意に決定できるが、<指向・偏向>の場合は、決定できない . なお、第 2 の例は、次項で示す

表 3 意味属性による名詞の訳し分けの可能性分類
Table 3 Capability of Differentiation of Noun Translation by Semantic Attributes

分類	分類	割合	名詞の例		
			見出し	意味属性	英訳語
1	訳し分け可能	55%	貝	<魚> <魚介類>	shellfish
				<殻>	shell
			スキー	<スポーツ>	skiing
				<遊び道具・運道具>	ski
2	場合により可能	24%	委員	<成員> <複数>	committee
				<成員> <単数>	member of a committee
			脱線	<指向・偏向> <目的>	deviation 方針や基準からの～
				<指向・偏向> <話>	digression 話の～
				<事件>	derailment 電車などの～
3	絞り込み可能	13%	木	<樹木>	tree 樹木
				<樹木>	shrub 灌木
				<材木>	wood 材木
				<材木>	lumber 製材
				<材木>	log 丸太
			麻	<作物 繊維>	flax 亜麻とその繊維
				<作物 繊維>	hemp 麻, 大麻とその繊維
				<糸・布>	linen 麻製品
				4	訳し分け不可能
<牙>	fang 犬や猫				
大学	<学校> <公共機関>	colleg 分科, 専科			
	<学校> <公共機関>	university 総合			
	<学校> <公共機関>	institute 理工学			

「絞り込み可能」と類似しているが、どの訳語に対しても、それに対して一意に決定できる意味属性が存在するので、この範疇に分類する。

<分類 3> 絞り込み可能

意味属性によって、訳語を一意に決定することはできないが、訳語候補の数を減らすことができる名詞を「絞り込み可能」とする。

表 3 の名詞「木」の例では、5 つある訳語候補が、<樹木> という意味属性を持つ語と <材木> という意味属性を持つ語に分けられる。意味属性が <樹木> の場合は、訳語は、"tree", "shrub" の 2 つに絞られるのに対して、意味属性が <材木> の場合は、"wood", "lumber", "log" の 3 つに絞られる。

また、名詞「麻」では、日本語文の解析によって意味属性が <糸・布> に決定した場合は、"linen" が選ばれるが、<作物> または <繊維> のいずれかに決定した場合は、"flax", "hemp" の 2 候補がともに同じ意味属性を持つため、どちらか一方を選択することはできない。しかし、この場合、訳語候補の数は、3 から 2 に減少したといえるので、このような名詞も「絞り込み可能」に分類する。

<分類 4> 訳し分け不可能

英訳語に付与された意味属性が全て同じ場合は、意味属性を訳し分けのための情報として利用できない。このような名詞を「訳し分け不可能」とする。

表 3 の名詞「牙」は、英語では象などの牙を "tusk"、犬や狼の牙を "fang" として使い分けるが和英辞書では、どちらの訳語も意味属性は <牙> であるため、両者を区別することはできない。また「学校」も同様である。

3.3 意味的包含関係による訳語制約効果

前節の結果から、訳語多義のない語を含め、各分類に属す語数の割合を図 1 に示す。

この図から、以下のことが分かる。

- (1) 訳語多義語のうち、ほぼ、半数に当たる 55% の名詞は、一意に訳し分けられる可能性がある。
- (2) 一意に訳し分けはできないが訳語候補の絞り込みに効果のある名詞は、「場合により可能」を含めて 32% である。

訳語多義語のうち、訳し分けで意味属性の効果が期待できる名詞は、上記の (1),(2) を含めて 87% にのぼる。効果が期待できない名詞は、13% にとどまっていることから、意味属性は大半の訳語多義語に対して有効であると言える。

次に、IPAL 基本名詞に対して、意味属性を使用した場合と使用しない場合の訳語選択の正解率を表 4 に示す。

この表から以下のことが推定される。

- (1) 訳語多義語の平均訳語数は、意味属性による訳語の選択を行う事で、おおそ半分減少する
- (2) 訳語多義語に対する訳語選択の正解率は、約 2

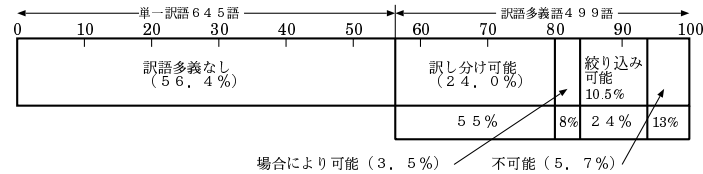


図 1 訳語選択から見た IPAL 基本名詞 1,144 語の分類

Fig. 1 Classification of IPAL Basic Nouns based on Selective Translations

表 4 意味属性による訳語制約効果の推定

Table 4 Effects of the Constrain by Semantic Attributes for Translation Word Selection

分類	評価項目	意味属性を	
		使用しない	使用する
訳語多義語 (499 語)	平均多義数	3.02	1.74
	正解率	38.7%	78.6%
基本名詞 (1,144 語)	平均多義数	1.88	1.32
	正解率	73.3%	90.7%

倍に向上する

- (3) その結果、基本名詞全体の訳語選択の正解率は、90% 程度まで向上する

以上から、名詞の訳語選択において、「意味属性」の情報がかなり有効であると期待される。

4. 例題検討 (新聞記事の例)

4.1 新聞記事に現れた名詞の訳語多義

前章では、IPAL 名詞の一語一語について (出現頻度を考慮しない場合について)、意味属性を使用したときの訳し分けの可能性について検討した。本章では、新聞記事を対象に、名詞が実際の文章で使用される頻度を考慮した場合について検討する。

(1) 新聞記事における IPAL 基本名詞の出現頻度

まず、新聞記事の対訳コーパス 1 万文を対象に、IPAL 基本名詞 1,144 語の使用頻度を調べた。この対訳コーパスは、朝日新聞、毎日新聞、読売新聞の政治欄、経済欄、社説、投稿などの記事を 10 グループに分けて、それぞれ 1,000 文ずつ集め、翻訳家によって英訳文を作成したもので、標準的な新聞記事文の集合と言える。

対象とした基本名詞のうち、使用されている名詞は 750 語で、総出現回数は 19,238 回であった。最も出現頻度が高い語は「問題」(424 回)で、以下「会社」(384 回)、「社長」(376 回)と続き、出現頻度の低い名詞ほど数が多くなっている。

(2) 新聞記事に現れた IPAL 基本名詞の多義

対訳コーパスで使用された基本名詞 750 語のうち、使用頻度が 50 回以上の名詞は、106 語であった。これらの名詞の累積度数は 13,069 回で、IPAL 基本名

詞の出現回数全体のほぼ 2/3 に相当する。そこで、これらの名詞の訳し方について調査分類した。その結果を表 5 に示す。また、各分類について、語数で見たときと累積出現度数で見たときの割合を図 2 に示す。

以下、分類された名詞の種類とそれらの訳し分けの方法について考察する。

< 分類 1 > 単一の訳語が使用される名詞

106 語中、単一の訳語が使用されている名詞は「東京」と「日本」の 2 語であった。いずれも地名を表す固有名詞である。このうち「日本」は、下記の例のように訳されないものもあったが、この分類に含めた。

例文 1) イギリス、フランス、ドイツを中心に日本から輸出し、欧州全体で月間三万台の販売を目指す。

They will export the VTR cameras to England, France and Germany, and they are expecting the sales of thirty thousand per month in the whole Europe.

日本語解析では、固有名詞について、例えば「平野」(ひらの<人名>、へいや<普通名詞>)、「国立」(くにたち<地名>、こくりつ<普通名詞>)などのように、人名、地名、一般語を識別する問題があるが、固有名詞であることが分かれば、2 章で述べたように訳語選択の問題は解決する。

< 分類 2 > ほぼ決まった訳語が使われる名詞

複数の訳語への訳し分けはあるものの、80%以上の割合で同一の訳語に訳される名詞を第 2 の分類とした。表 5 の例では「学校」の英訳語としては、「school」、「institute」、「academy」、「academic」などが使われているが、これらのうち、「school」が全体の 91.6%を占めている。従って、新聞記事の場合、訳語をあらかじめ「school」と決めておけば、9 割の正解率が得られる。

< 分類 3 > 有力な候補がある名詞

40%を越える確率で選択される訳語を持つ名詞があり、他の訳語の選択される確率の低い名詞を第 3 の分類とした。表 5 の例では「計画」に対して訳語「plan」が選択される割合が、61.9%であるから、最低限 6 割の正解率が保証できる。

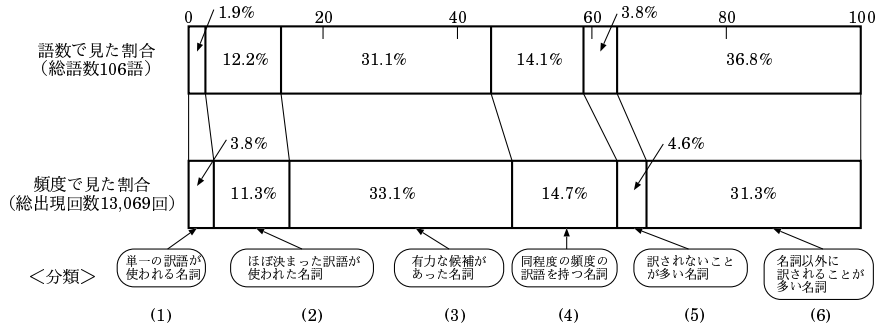


図 2 訳語選択から見た名詞の分類 (対象新聞記事)
Fig. 2 Classification of Nouns for Translation Words Selection

表 5 訳語選択から見た名詞の分類
Table 5 Classification of Nouns for Translation Words Selection

#	分類	該当する名詞 (出現回数)	訳語分布の例		
			見出	訳語	度数 (割合)
1	単一の訳語が使われた名詞	日本 (361), 東京 (255) (計 2 語)	日本	Japan なし	354(98.1%) 7(1.9%)
2	ほぼ決まった訳語が使われた名詞	銀行 (294), 世界 (239), 改革 (219), 工場 (105), 価格 (103), グループ (81), 子供 (69), 土地 (68), 他 (65), 領土 (63), 学校 (60), 緊急 (58), 条件 (63) (計 13 語)	学校	school academy institute academic	55(91.6%) 2(3.3%) 1(1.7%) 1(1.7%)
3	有力な候補がある名詞	問題 (424), 会社 (384), 生産 (301), 社会 (222), 計画 (210), 私 (198), 投資 (159), 協力 (159), 国民 (153), 経営 (136), 率 (135), 影響 (121), 営業 (115), アジア (115), 建設 (107), 人 (105), 規模 (96), 組織 (92), 代表 (90), 都市 (88), 援助 (87), 野党 (84), 報告 (83), 中央 (83), 地方 (76), 焦点 (75), 提案 (71), 基本 (68), 電気 (68), 批判 (60), 団体 (58), 立場 (52), 段階 (51) (計 33 語)	計画	plan(n) plan(v) project program scheme schedule	130(61.9%) 30(14.2%) 9(4.3%) 8(3.8%) 2(1.9%) 1(0.5%)
4	同程度の頻度の訳語を持つ名詞	企業 (354), 拡大 (247), 機関 (202), 関係 (192), 時間 (165), 管理 (120), 利用 (116) 国 (103), 設備 (83), 西 (78), 課題 (58), 能力 (56), 議長 (56), 目的 (50), 幹部 (50) (計 15 語)	能力	ability capacity capability	17(30.3%) 17(30.3%) 16(28.5%)
5	訳されないことが多い名詞	社長 (354), 本社 (120), 形 (89), 方向 (51) (計 4 語)	方向	なし direction way course	33(64.7%) 13(25.4%) 4(7.8%) 1(2.0%)
6	名詞以外に訳されることが多い名詞	明らか (232), 中 (219), 対応 (218), 強化 (187), 額 (186), 中心 (176), 発表 (164), 量 (156), 検討 (143), 以上 (121), 前 (117), 指導 (115), 実現 (112), 期待 (110), 開始 (109), 展開 (106), 決定 (103), 実施 (101), 海外 (100), 予想 (86), 低下 (82), 程度 (81), 可能 (79), 例 (71), 提出 (69), 内容 (68), 先 (67), 背景 (67), 間 (67), 採用 (66), 内 (65), 外国 (64), 上昇 (63), 筋 (59), 進出 (56), 確認 (56), 禁止 (54), 後 (51), 手 (50) (計 39 語)	開始	start begin open commence <名詞>	43(39.4%) 28(25.7%) 7(6.4%) 2(1.8%) 29(26.6%)

<分類 4> 同程度の頻度の訳語を持つ名詞

突出して選択される訳語が無く、同程度の頻度の複数の訳語を持つ名詞を分類 4 とした。表 5 の例では、名詞「能力」に対して、3 つ訳語, "ability", "capacity", "capability" が、ほぼ、均等な割合で選択されている。

<分類 5> 訳されないことが多い名詞

日本文の中には現れるが、英訳文では現れないことの名詞 4 語が該当する。これらの 4 語は、用法から見て 2 種類に分類される。

まず、「方向」、「形」の 2 語は、多くの場合、特定の言い回しの中で用いられている。例えば、下記の例文では、名詞「形」が「～の形で」のように連語の一部として使用されるとき直接対応する訳語はない。このような連語的な表現はあらかじめ収集しておき、名詞

の訳語選択処理の対象外とするのが適切と考えられる。

例文 2) 出資額は明らかにしていないが、口社の増資分を引き取る形で三〇%程度の出資比率となる見込み。

Although the investment amount has not been made clear, they will take over the increased capital to be paid in by Robotec. Therefore, their total capital investment will be somewhat around 30 percents.

一方、残りの「本社」、「社長」の 2 語は、英文中、訳語が存在する場合もあったが、多くは、下記の例に示すように、記事中の人名や社名の補足説明に使用されており、省略される場合の方が圧倒的に多い語である。但し、これは、新聞記事の特殊性によるもので、一般性に乏しいと思われる。

例文 3) 電子部品メーカーのコパル電子(本社 東京, 社長 山田康弘氏, 資本金十二億一千六百十五万円)は電子センサー事業を拡大する。

The electronic parts maker Coparu Electronics will expand its electron sensor business.

<分類 6> 名詞以外に訳されることが多い名詞

名詞以外の英訳語に訳されることが多い名詞を言う。例えば, 表 5 の「開始」というサ変名詞は動詞として訳されることが多く, 名詞にはあまり訳されていない。

このほかにも, 副詞や前置詞に訳されることが多い「前」や「中」などの位置, 時間, 程度等を表す名詞や「手を結ぶ」など特定の言い回しでよく用いられる名詞は, 第 6 の分類に入れた。下記の例文 4) では「前」という名詞が, 副詞“before”に訳されている。また, 例文 5) では名詞「手」は, 単独では訳されず「手に入れる」として“get”に訳されている。

例文 4) 一年前に発表した「システム/390」の全容が整ったことになる。

This means that the whole picture of the “System/390” announced one year before has now been realized.

例文 5) 一〇%, 二〇%の下落では一般サラリーマンが手に入れることは, 夢のまた夢だ。

At a drop of 10% or 20%, for the ordinary salaried worker getting a piece of land is only the dream of a dream.

これらの名詞は, 前後にある他の語との組合わせで訳語が決まる場合が多い。日本語名詞から英語名詞へのタイプの選択ではなく, 別の観点から解決を図るのが適当と考えられる。

例えば, サ変名詞を動詞として訳す場合は, 従来の動詞の訳語選択の方法が適用できると期待される。また, 時間, 位置, 程度などを表す「中」「先」などの名詞を副詞や前置詞に訳す場合については, その種類もある程度限られているので, 個別の翻訳規則を作成することが期待される。また, ほかに「量」「額」など「情報量」「投資額」などのように複合名詞の一部として使用される名詞もあるが, これらは複合語全体の翻訳技術の中で検討すべき課題と考えられる。

4.2 意味的用法による訳語制約効果

現状の翻訳システムでは, 名詞の訳語を選択する場合, 最も出現頻度の高い訳語に訳す方法が一般的である。分類 1, 分類 2 の 15 語は, この方法で 90%以上の正解率が期待できるが, 分類 3 と分類 4 の 48 語は, この方法は適切と言えない。中でも, 分類 4 「同程度の頻度の訳語を持つ名詞」の訳語選択は難しい。そこ

で, 分類 4 に含まれる 15 語を対象に意味属性を使用した場合の訳語選択の精度について評価した。

その結果によれば, 和英辞書で見たこれらの語の平均多義数は, 3.0 であるのに対して, 意味属性を使用した場合, 平均多義数は 1.4 に減少し, 訳語正解率は, 53.8%から 92.2%に向上する見込みとなっている。

5. 訳語選択に必要な知識について

前章までの結果に基づき, 名詞の訳語選択から見た名詞の意味属性体系の拡張性とその他の知識の必要性について検討し, 今後の課題について考察する。

5.1 意味属性による方法の拡張性

第 4 章では, 名詞の「意味的用法」に関する知識として「意味属性体系」を使用した。これは, 動詞の訳語選択を目的として開発されたものであった。ここで, 名詞の訳語選択を対象とする場合は, この体系をどのように改良することができるか, また改良の効果はどれくらいかについて考える。

複数の訳語を持つ基本名詞 499 語のうち, 一意に訳語が決定できなかった 224 語を対象に, それぞれの訳語候補の違いを調べた。それを基に現在の意味属性を修正, 拡張することで解決が見込まれる項目, および, 英語訳語に付与された意味属性の見直しで改良できる見込みの項目を抽出した。その結果を表 6 に示す。

例えば, 表中の「単数と複数」を区別する例では, 「さくら」の訳語候補に付与されている<補佐>という意味属性を「単数」「複数」に分離し, “claqueur”の意味属性を<補佐・単数>, “claque”の意味属性を<補佐・複数>とすれば, 2 つの訳語候補の訳し分けが可能となることを意味する。

表 6 の方法の適用対象となる名詞を調べた結果では, 問題とされる 224 語のうち, 訳語が一意に決定できる名詞は 30 語と推定される。これは, 複数の訳語候補を持つ見出し語全体 499 語から見ると 6%で, あまり大きな効果は期待できない。

5.2 新しい視点による知識導入の必要性

前節の結果から, 訳語選択の精度をさらに向上させるためには, 名詞の「意味的用法」に関する分類知識以外の知識を併用することが必要と考えられる。そこで, ほかにどのような知識が必要であるかについて検討した結果を表 7 に示す。

表 7 において, 例えば, 第 1 行目の見出し語「えさ」の訳し分けでは「何のための餌か」に関する知識が必要であることを示す。また, 第 2 行目の「傷」の場合は「何が原因でできた傷か」の知識が必要である。

表から分かるとおり, これらの名詞は, 日英両言語

表 6 名詞の意味属性体系の改良が期待できそうな項目
Table 6 Expectation of the Improvements of Semantic Attribute System

分類	訳し分け可能となる名詞の例	
単数と複数	「さくら」	劇場の(その中の一人): claqueur <補佐> 劇場の(集会的に): claque <補佐>
一般と特定	「校舎」	学校の建物: school building <家屋(本体)> <学校> 特に小学校: schoolhouse <家屋(本体)> <学校>
具体と抽象	「腕」	具体物: arm <腕> 抽象物(技術): skill <腕>
詳細分類	「えび」	車エビ: prawn <えび かに たこ いか 魚介類> 小エビ: shrimp <えび かに たこ いか 魚介類> 伊勢エビ: lobster <えび かに たこ いか 魚介類>
男と女	「牛」	(雄): bull <獣> <男> (雌): cow <獣> <女>
自然物と化工品	「汁」	果物, 野菜, 肉など: juice <汁> <液体(その他)> 吸い物: soup <汁> <コーヒー ジュース>
英語と米語	「種」	桃など: pit <食品> (米) 梅など: stone <食品> (英)
口語と文語	「自転車」	口語的: bike <乗り物> <スポーツ> 文語的: bicycle <乗り物> <スポーツ>

表 7 意味的包含関係以外の知識を必要とする例
Table 7 Required Knowledge except for Semantic Inclusion

知識種別	見出	意味属性	英訳語
目的	えさ	<飼料>	feed 動物などの「餌」の場合 bait 魚, 動物などの「餌」の場合
原因	傷	<怪我>	cut 切り傷の場合 wound 物理的, 精神的な傷の場合 injury 偶然に受けた傷の場合
所有者	爪	<爪>	claw 鳥や獣の場合 nail 人の場合
視点	表	<表>	face 裏の反対の場合 surface 内面の反対の場合
形状	帽子	<帽子>	hat 縁つき「帽子」の場合 cap 縁なし「帽子」の場合
部位	髭	<ひげ>	beard 「あごひげ」の場合 mustache 「口ひげ」の場合 whiskers 「頬髭」の場合
対象部分	顔	<顔>	head 首から上 face 顔面

間で表す対象の範囲が微妙にずれているものが多く、そのずれ方はさまざまであるため、それぞれの観点から訳し分けに必要な知識を体系的に準備することは容易でない。しかし、前章までの検討から、このような訳し分けの必要な名詞は、限られていると見られるので、個別的に翻訳規則を検討することが可能と思われる。

6. あとがき

日英機械翻訳における名詞の訳語選択の問題を解決するための第1ステップとして、日本語の基本的な名詞に対する英訳語の多義構造について調査すると共に、名詞の「意味的用法」の知識に着目した英訳語選択の可能性について検討した。

具体的には、IPAL 辞書に収録された計算機用日本

語基本名詞に対して、和英辞書を使用して名詞の持つ英訳語数の分布を調べ、そのうち、複数の英訳語を持つ名詞を対象に、名詞の「意味的用法」として「日本語語彙大系」で定義された「一般名詞意味属性体系」を使用した場合の訳語選択の精度を調べた。その結果によれば、基本名詞 1,144 語の中で、複数の訳語を持つ名詞は約 5 割 (499 語) で、そのうち、55% は、「意味属性」の知識によって一意に訳語が決定できること、また、それを含む 87% は、訳語候補が絞り込める可能性のあることが分かった。

また、新聞記事 1 万文を対象に使用頻度の高い基本名詞 750 語の検討結果では、使用頻度の高い基本名詞 106 語の平均多義数は 3.02 から 1.74 に減少し、訳語正解率は、38.7% から 78.6% に向上する可能性があることが分かった。特に、同程度の頻度の複数の訳語を持つ名詞の場合は、平均多義数が 3.0 から 1.4 に減少し、訳語正解率は、53.8% から 92.2% に向上する見込みである。

訳語選択に失敗した用例の分析では、「意味属性体系」の改良で約 6% の精度向上が見込まれるが、それ以上の改良は困難であることから名詞の「意味的用法」の知識を使用した方法の限界は、上記の精度+6%程度と予想される。なお、「意味的用法」の分類知識で訳し分け困難な名詞の数は、かなり限定的と見られるので、今後、個別に検討することが期待される。

参考文献

- 1) Allen, James, "Natural Language Understanding Second Edition", The Benjamin/

- Cummings Publishing Company, Inc., 1995
- 2) 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
 - 3) 小嶋秀樹, 伊藤昭, "辞書にもとづいて語彙をクラスタリングする試み", 言語処理学会第1回年次大会論文集, pp.205-208, 1995
 - 4) Brown P., Della, S., Della, V., Mercer, R., "Word-sense disambiguation using statistical method", Annual Meeting of ACL, pp.264-270, 1991
 - 5) Yarowsky, D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpra", COLING-92, pp.454-460, 1992
 - 6) Dagan, I., Itai, A., "Word-sense disambiguation using a second language monolingual corpus", Computational Linguistics, Vol.20, No.4, pp.563-596, 1994
 - 7) Agirre, E. and Riguu, G., "Word Sense Disambiguation using Conceptual Density", COLING'96, pp.16-22, 1996
 - 8) Li, H. and Abe, N., "Word Clustering and Disambiguation Based on Co-occurrence", COLING-ACL'98, pp.749-755, 1998
 - 9) Barriere, C., "Redundancy: helping semantic disambiguation", COLING-ACL'98, pp.103-109, 1988
 - 10) Wilks, Y. and Stevenson, M., "Word Sense Disambiguation using Optimised Combinations of Knowledge Sources", COLING-ACL'98, pp.1398-1402, 1998
 - 11) 野美山浩, "目的言語の知識を用いた訳語選択とその学習性", 情報処理学会自然言語処理研究会, 91-NL-86, 1991
 - 12) 鈴木, 太細, "日英機械翻訳における共起表現の扱い", 情報処理学会自然言語処理研究会, 91-NL-82, 1991
 - 13) Niwa, Y. and Nitta, Y., "Co-occurrence vectors from corpra vs. distance vectors from distances", COLING'94, pp.304-309, 1994
 - 14) 福本文代, 辻井潤一, "コーパスに基づく動詞の多義解消", 自然言語処理, Vol.4, No.2, pp.21-39, 1997
 - 15) 平岡冠二, 松本裕二, "共起情報を用いた多義動詞の類別と名詞のクラスタリング", 言語処理学会第1回年次大会論文集, pp.149-152, 1995
 - 16) Fujii, A., Inui, K., Tokunaga, T., Tanaka, H., "Case Contribution in Example-Based Verb Sense Disambiguation", 自然言語処理, Vol.4, No.2, pp.111-123, 1997
 - 17) 内田将夫, 板橋秀一, "シソーラス上に動的に構成される標本空間における動詞の多義解消", 自然言語処理, Vol.4, No.2, pp.27-50, 1997
 - 18) 柏野和佳子, "解析と生成のための共起情報の記述方法", 自然言語処理学会第五回年次大会論文集, pp.205-208, 1999
 - 19) 宇津呂武仁, 松本裕二, 長尾真, "2言語対訳コーパスからの動詞の格フレーム獲得", 情報処理学会論文誌, Vol.34, No.4, pp.913-924, 1993
 - 20) Grishman, Ralph, "Generalizing automatically generated selectional patterns", COLING-94, pp.742-747, 1994
 - 21) 田中英輝, "動詞訳語選択のための「格フレーム木」の統計的な学習", 自然言語処理, Vol. 2, No.3, pp.49-72, 1995
 - 22) 秋葉泰弘, 石井恵, 金田重郎, "人手作成ルールと事例に基づく英語動詞選択ルールの学習", 自然言語処理, Vol.3, No.3, pp.53-68, 1996
 - 23) 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦, "日本語語彙大系 1. 意味体系", 岩波書店, 1997
 - 24) 白井諭, 井上浩子, 横尾昭男, 池原悟, "日英機械翻訳における用言の訳し分けと構文意味辞書", 言語処理学会第1回年次大会論文集, pp.265-268, 1995
 - 25) 角田達彦, 田中英彦, "英語名詞の多義性解消における文脈としての場面情報の評価", 自然言語処理, Vol.3, No.1, pp.4-27, 1996
 - 26) 北村, 荻野, "日英翻訳における連体修飾句の訳し分け", 情報処理学会自然言語処理研究会, 90-NL-75, 1990
 - 27) 桑畑和佳子, 本多, "IPAL 名詞辞書における多義構造の記述", 第16回 IPA 技術発表会, pp.189-200, 1997
 - 28) 桑畑和佳子, 橋本美奈子, 青山文啓, "IPAL 名詞辞書による多義性解消のためのコーポレーションの分析", 情報処理学会論文誌, Vol.39, No.6, pp.1925-1934, 1998
 - 29) 池原悟, 村上仁一, 車井登, "日英機械翻訳のための日本語抽象名詞の文法的・意味的用法の分類", 自然言語処理, Vol.9, No.1, pp.117-134, 2001
 - 30) IPAL「計算機用日本語基本名詞辞書 IPAL 解説編」, 情報処理振興事業協会技術センター, 1996
 - 31) 池原悟, 宮崎正弘, 白井諭, 林良彦, "言語における話者の認識と多段翻訳方式", 情報処理学会論文誌, Vol.28, No.12, pp.1269-1279, 1987
 - 32) 池原悟, 宮崎正弘, 横尾昭男, "日英機械翻訳のための意味解析用の言語知識とその分解能", 情報処理学会論文誌, Vol.34, No.8, pp.1692-1704, 1993
 - 33) 金出地真人, 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, 情報処理学会第63回全国大会, 6Y-04, 2-267-268(2001)

(平成 14 年 1 月 1 日受付)

(平成 14 年 1 月 1 日採録)

池原 悟 (正会員)

1967年大阪大学基礎工学部電気工学科卒業。1969年同大学院修士課程修了。同年日本電信電話公社に入社。数式処理, トラフィック理論, 自然言語処理の研究に従事。1996年スタンフォード大学客員教授。現在, 鳥取大学工学部教授。工学博士。1982年情報処理学会論文賞, 1993年同研究賞, 1995年日本科学技術情報センター賞(学術賞), 同年人工知能学会論文賞, 2002年電気通信普及財団賞(テレコム・システム技術賞)受賞。電子情報通信学会, 人工知能学会, 言語処理学会, 機械翻訳協会各会員。

村上 仁一

1984年筑波大学第3学群基礎工学類卒。1986年筑波大学修士課程理工学研究科理工学専攻修了。1986年NTTに入社。NTT情報通信処理研究所に勤務。1991年国際通信基礎研究所(ATR)自動翻訳電話研究所に出向。1997年鳥取大学工学部知能情報工学科に転職。現在に至る。主に音声認識のための言語処理の研究に従事。電子通信情報処理学会, 日本音響学会, 言語処理学会, 各会員。

桐澤 洋

1998年鳥取大学工学部知能情報工学科卒業。2000年大学院修士課程修了。同年株式会社メイテックに入社。現在に至る。