日英機械翻訳における名詞の訳語選択

桐澤 洋　池原 悟　村上仁一
鳥取大学大学院工学研究科
{kirisawa,ikehara,murakami}@ike.tottori-u.ac.jp

1 はじめに

機械翻訳を行う際の問題点の1つとして、多義を持つ単語の訳語選択がある。ある単語に対して複数の訳語が存在する場合、文として正しい翻訳結果を得るためには、その中から適切な訳語を選びだす必要がある。これまでの訳語選択法の研究としては、格の情報と意味属性を用いた方法(北村、荻野 1990)などが提案されているが、名詞自体に訳し分けの情報を求める研究や、名詞一般についての分析を行った論文はあまり見受けられない。名詞の多義構造についての論文としては(桑畑、本多 1997)があるが、日本語に閉じた研究であり英語との対応関係は不明である。これは、名詞は数が多い上に意味が多様で、今まで有効な手掛かりが無かったためである。本稿ではその手掛かりを得るために以下の2つについて検討する。まず、意味属性体系の持つ語義識別能力に着目し、この能力を用いることで名詞の訳し分けがどの程度可能となるかを明らかにするとともに、日本語名詞と英訳語の多義数の関係を調べる。次に、対訳コーパスを用いた統計を行い、実際の文章における日本語と英語の対応関係を探る。

2 意味属性の語義識別能力

2.1 意味属性を用いた訳語選択

「意味属性」とは、「ある単語が意味的にどんな使われ方をするかという意味的用法を整理し、体系化したもの」(池原 他 1997)である。単語の「意味的用法」は単語の語義から派生することを考えると、実際に使用された文中での単語の「意味的用法」がわかれば、その単語がどの語義で使用されたか判断できる可能性がある。従って、意味属性は訳語選択に役立つと期待できる。

例えば、「犬」という日本語名詞には"dog"、"spy"という2つの英訳語があり、それぞれ[獣]、[スパイ]という意味属性を持つ。そこで「犬」という単語が使われている日本語文を解析し、その単語の意味属性が決定できれば対応する英訳語を決定できる。つまり、「犬」の意味属性が[獣]と決まれば"dog"が、[スパイ]と決まれば"spy"が選択される。一方、「えさ」という日本語名詞の英訳語"feed"、"bait"はどちらも[飼料]という意味属性を持つ。これは、日本語では同義として扱うが英語では使い分けがあるために複数の訳語が対応する名詞であり、日本語から見た場合「意味的用法」に違いが無いため意味属性による訳し分けはできない。

なお、一般に意味属性は1つの単語に対して複数付与されているが、今回の研究では日本語文の解析によって一意に決まると仮定して検討を進める。

2.2 検討対象

意味属性についての検討では、日本語内での語義分類と日英を対比した場合の語義分類の違いを調べるとともに、単語意味属性の持つ語義識別能力を明らかにするため、以下の辞書を使用する。

(1) 計算機用日本語基本名詞辞書 IPAL

通産省の外郭団体である情報処理振興事業協会が作成した日本語の辞書であり、複数の言語学者により選定された日本語の基本名詞1,081語が収録されている。それぞれの名詞は仮名表記を見出し語とし、複数の日本語表記や語義など、見出し語ごとに詳細な情報が記述されている。

(2)ALT-J/E 日英対照一般名詞辞書

NTTが作成した日英対照辞書。約60,000語の見出し語が収録されており、それぞれの英訳語のもつ意味属性が辞書の情報として記載されている。なお、この辞書はNTTが開発した「ALT-J/E」という機械翻訳システムで実際に用いられている。

(3) 日本語語彙大系

岩波書店より出版されている全5巻からなる辞書で、日本語意味解析のための単語体系や構文体系などが収録されている。今回の検討では、この中でも「第1巻 意

味体系」に収録されている一般名詞意味属性を用いる。これは一般名詞に対して[具体]、[関係]など約2,700の属性に分け、木構造としてまとめたもので、約30万語の名詞の意味的用法が単語意味属性を用いて定義されている。

検討は、これらの辞書に収録されている名詞のうち、IPALの辞書に登録されている日本語の基本名詞1,081語を対象に行う。名詞はよく使われるものほど多様な意味を持っていると考えられるが、IPALの辞書に収録されている基本名詞を対象とすることで特に多義の多い名詞を中心に検討を進めることができ、名詞の多義性解消の検討として十分な結果が得られると期待できる。

2.3 辞書上で見た語義分解能

2.3.1 IPALとALTの対応付け及び比較

まず双方の辞書に収録されている見出し語を日本語表記をもとに対応させて対応表を作成し、この表を元に日本語の中で見た場合と、英語と対比して見た場合とでの語義分類の違いと、多義数の関係を調べた。

その結果、IPALに収録されている1,081語のうち94%にあたる1,014語に対してALTの1,144語を対応づけることができた。このことからALTの辞書が日本語の基本名詞を十分に収録しているといえる。また、双方の多義数を比較した結果を表1、図1に示す。これらを見ると、名詞の多義数は日英を対比して見た場合よりも日本語の中で見た場合の方が多いことが分かる。

表1 IPALとALTの比較

	IPAL	ALT
平均多義数	2.13	1.88
最大多義数	18	12

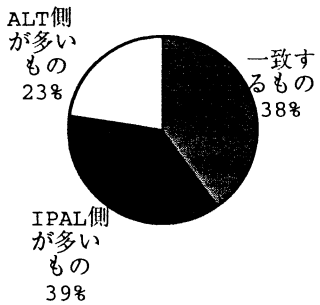

図1 IPALとALTの多義数の関係

2.3.2 意味属性の語義識別能力の検討

次に、対応表から2つ以上の英訳語を持つ見出し語を全て抽出し、意味属性による訳し分けがどの程度可能かを検討した。

対象となった499語を調べたところ、表2のような結果が得られた。それぞれの例も同時に示す。この表より、87%の名詞に対して何らかの効果が得られたことが分かる。また、意味属性を用いることで平均多義数は3.02から約半分の1.74に減少、正解が得られる確率は38.7%からおよそ2倍の78.6%に向上した。以上より、意味属性は訳語選択において有効であると言える。

表2 意味属性の語義識別能力とその例

	見出し語	意味属性	英訳語	
訳し分け可能 55%	いなか	[村落]	都会に対して	:country
		[郷里]	故郷	:home
絞り込み可能 24%	木	[樹木]	樹木	:tree
		[樹木]	灌木	:shrub
		[材木]	材木	:wood
		[材木]	製材した	:lumber
		[材木]	丸太	:log
場合により訳し分け可能 8%	あわれ	[同情] [趣き]	哀れみ	:pity
		[悲しみ] [趣き]	悲しみ	:sadness
訳し分け不可能 13%	牙	[牙]	象などの	:tusk
		[牙]	犬・おおかみの	:fang

3 新聞記事における名詞の訳語選択

3.1 名詞の出現頻度

統計によって実際の文章の中での名詞の出現頻度を探り、出現頻度を考慮にいれた訳語選択について検討する。データベースには新聞記事の対訳コーパス10,000文を用いる。これは様々な日本語の新聞から記事を集め、翻訳家によって英訳文を作成したもので、記事内容は「政治」や「経済」から「読者投稿」まで広く含む。また、対象とする名詞は2章で用いたIPALの辞書に収録されている日本語基本名詞1,081語に限る。

統計をとった結果は図2のようになった。1,081語の総出現回数は19,238回で、最も出現頻度が高いのは「問題」の424回。以下、出現頻度の低い名詞ほど数が多くなっている。

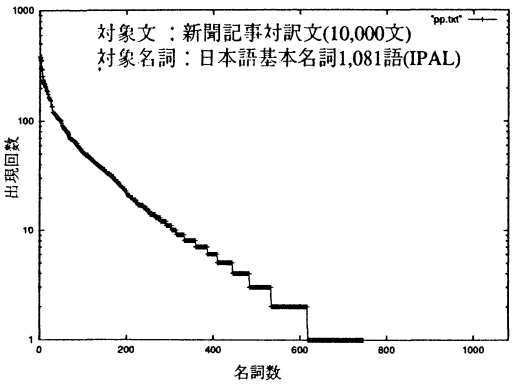

図 2 IPAL 名詞の出現回数の分布

3.2 訳語候補

次に、出現頻度が 50 回以上の見出し語 106 語 (累積度数:12,744) を対象に、各見出し語ごとに対応する英語表現を調べ、訳語の出現する割合によって表 3 の通り、6 つに分類した。

表 3 英訳語の出現頻度による名詞の分類

1) 多義のない名詞
「東京」「日本」 (計 2 語)
2) 訳語がほぼ一意に決定できる名詞
「条件」「緊急」「学校」「領土」「他」「土地」「子供」 「グループ」「価格」「工場」「改革」「世界」「銀行」 (計 13 語)
3) 有力な候補がある名詞
「段階」「立場」「団体」「批判」「電気」「基本」「提案」 「焦点」「地方」「中央」「報告」「野党」「都市」「援助」 「代表」「組織」「規模」「人、ひと」「建設」「アジア」 「営業」「影響」 「率」「経営」「国民」「協力」「投資」 「私」「計画」「社会」「生産」「会社」「問題」 (計 33 語)
4) 同程度の出現頻度の候補を持つ名詞
「幹部」「目的」「議長」「能力」「課題」「西」「設備」 「国、故郷」「利用」「管理」「時間」「関係」「機関」 「拡大」「企業」 (計 15 語)
5) 訳されないことが多い名詞
「方向」「形」「本社」「社長」 (計 4 語)
6) 名詞以外に訳されることが多い名詞
「手」「後、あと」「禁止」「確認」「進出」「筋」「上昇」 「外国」「内、うち」「採用」「間」「背景」「先」「内容」 「提出」「例」「可能」「程度」「低下」「予想」「海外」 「実施」「決定」「展開」「開始」「期待」「実現」「指導」 「前」「以上」「検討」「量」「発表」「中心」「明らか」 「額」「強化」「対応」「中」 (計 39 語)

1) 多義のない名詞

日本語と英語が一対一に対応する名詞として、地名である「東京」「日本」の 2 語が挙った。それぞれ"Tokyo"、"Japan" と訳されるか、全く訳されないかのいずれかであった。

2) 訳語がほぼ一意に決定できる名詞

英訳語の中に出現頻度の割合が 80%以上を占める候補がある名詞をここに分類した。例えば、表 4 に挙げた「学校」という名詞の英訳語としては、"school", "institute","academy", "academic" などが使われていたが、その中で"school" が全体の 91.6%を占めているため、これを訳語として選択することで 9 割の正解率が得られることになる。

3) 有力な候補がある名詞

英訳語の中に 40%を越える候補が 1 つだけあり、他の候補は出現頻度の低い名詞を「有力な候補がある名詞」とした。表 4 の「計画」の例では、名詞である"plan" の出現頻度が 61.9%と他の訳語候補に比べて突出して高いため、これを最も適当な候補として選択することで 6 割の正解率を得ることができる。

4) 同程度の出現頻度の候補を持つ名詞

突出した候補がなく、同程度の出現頻度の候補を持つ名詞にあたる。表 4 の「能力」の項目をみると、3 つある訳語候補の出現頻度がそれぞれ 3 割程度で、頻度統計の結果を用いてこの中から 1 つを選びだすことはできない。

5) 訳されないことが多い名詞

日本語文の中には現れるが、英訳文では訳されていないことが多い名詞をここに分類した。たとえば以下の例文では、括弧書きで「コパル電子」についての補足説明が記述されているが、英訳文ではこの部分はまったく訳されていない。今回用いた対訳コーパスではこのような補足説明が省略されていることが多く、括弧書きの中でよく使われる「本社」や「社長」などはあまり訳されていない。

例) 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.

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の多い日本語名詞をここに分類した。例として、表4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などがここに分類された。

表4 英訳語の出現頻度による分類の例

		訳語候補	度数(割合)
ほぼ一意に決定できる名詞	学校	school	55(91.6%)
		institute	1
		academy	2
		academic	1
有力な候補がある名詞	計画	plan(n)	130(61.9%)
		plan(v)	30(14.2%)
		project	9
		program	8
		scheme	2
		schedule	1
同程度の出現頻度の候補を持つ名詞	能力	ability	17(30.3%)
		capacity	17(30.3%)
		capability	16(28.5%)
名詞以外に訳されることが多い名詞	開始	start	43(39.4%)
		begin	28(25.7%)
		open	7(6.4%)
		commence	2

3.3 訳語選択法の検討

3.2で挙げたの6つ分類について検討する。まず1)の「多義のない名詞」はもともと多義が無いため、2)の「ほぼ一意に決定できる名詞」は最もよく使われる候補を選ぶことで8割以上の正解率が得られるため、5)の「訳されないことが多い名詞」は省略されることが多いため、それぞれ大きな問題にはならないだろう。また3)の「有力な候補がある名詞」についても突出した候補があるため、それを選択することである程度の正解率を得ることができる。次に6)の「名詞以外に訳されることの多い名詞」ついては、他の語との組み合わせで訳語が決まるものが多く、名詞単独で問題になることはあまりないと思われる。最後に、分類4)の「同程度の出現頻度の候補を持つ名詞」が最も問題になると思われるが、これらに対して意味属性による訳し分けを行った結果、辞書上での平均多義数が3.0から1.4に減少、平均正解率が53.8%から92.2%に向上するという効果が得られた。しかし、意味属性の効果が全く無い名詞もあるため、これらの名詞については見出し語ごとに個別に検討する必要があるだろう。

なお、1)から4)の分類にあたる63語に対して最も使用頻度の高い候補を選ぶという方法を適用した場合、48.6%の平均正解率が得られた。

4 結論

本稿では名詞の訳語選択の問題に対して、単語意味属性体系の語義識別能力に着目し、その有効性を明らかにすると共に、対訳コーパスを用いた統計によって出現頻度を考慮に入れた訳語選択法について検討した。

その結果、意味属性を用いることで複数の訳語をもつ名詞の87%に対して訳し分けや候補の絞り込みなどの効果が得られたほか、平均多義数が約半分に減少、正解の得られる確率がおよそ2倍に向上するなどの効果もあった。また、意味属性では全く訳し分けることのできない名詞も13%ほどあったが、対訳コーパスによる統計の結果をみるとこれらの名詞は出現回数の少ないものが多く、新聞記事を対象とした訳し分けではそれほど問題にはならないだろう。

新聞記事を用いた統計では、訳し分けにおいて特に問題となりそうな名詞は、検討の対象とした106語のうち14%ほどであった。これらに対して、意味属性による訳し分けを行った結果効果は得られたものの、意味属性では訳し分けることのできない名詞も含まれていたため、各名詞ごとに個別ルールを作成した方が良いと思われる。

参考文献

池原 他 (1997)：日本語語彙大系 1. 意味体系、岩波書店

北野、荻野 (1990):日英翻訳における連帯修飾句の訳し分け、情報処理学会研究報告書,vol.90,No5、75-10

桐澤、池原 (1998):名詞多義構造の解析と訳語選択法、電子情報通信学会ソサイエティ大会,D-5-6

桐澤、池原、村上 (1999):名詞の訳語選択における意味属性の有効性、電子情報通信学会技術研究報告,vol.99,No88、NLC99-5

桑畑、本多 (1997):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