日中機械翻訳における名詞訳語の選択

展瑜　徳久雅人　池原悟　村上仁一
鳥取大学大学院工学研究科
{zhanyu, tokuhisa, ikehara, murakami } @ike.tottori-u.ac.jp

1 はじめに

機械翻訳では，動詞や名詞の訳語選択が大きな問題となっている．特に，名詞に複数の訳語が存在する場合，その中から適切な訳語を選び出すことが困難である．日英機械翻訳においては，「日本語語彙大系」[1]の「一般名詞意味属性」(以後，意味属性と略記する)と「結合価パターン」を用いることで約 90%の訳し分けが可能であることがわかっている[2]．しかしそれは，膠着言語族である日本語から屈折言語族である英語への翻訳の場合に有効性が確認されているものである．孤立言語族である中国語への翻訳の場合にも有効であるかは明らかではない．さらに「日本語語彙大系」は，元々日英翻訳のために，まとめられたものであり，文化や歴史的背景のことなる中国語にどれだけ流用できるのかは未知である．

そこで，本研究では日中機械翻訳において，意味属性を用いた日本語の基本名詞の訳語選択の可能性を検討する．

2 日本語基本名詞の訳語多義

2.1 調査対象

日本語の基本名詞に対して，中国語の訳語の持つ多義の量的構造を明らかにするため，「計算機用日本語基本名詞辞書 IPAL」[3](以後，IPAL 辞書と略記する)と「日中辞典」[4]を用いる．前者は情報処理振興事業協会が作成した日本語辞書であり日本語で頻繁に用いられる基本名詞が収録されている．後者は北京・対外経済貿易大学と北京・商務印書館及び小学館の共同編集によって作られた人間用の辞典である．

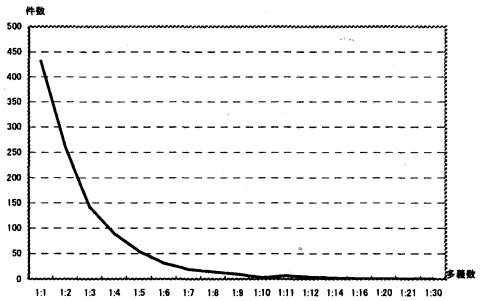
本研究では，IPAL 辞書に登録された日本語の基本名詞 1,081 語を対象に，「日中辞典」で定義された訳語の多義を調べる．なお，多義の判断は，原則として「日中辞典」における語義を基準とする．しかし，そこで分けられた訳語の中でも意味の差あるいは，使われ方の差が比較的大きい単語が含まれる場合，違う語義とした．そうでない場合は同義として1つの語義に属するものとした．

また，IPAL 辞書の中には，日本語の表記が異なるが，同一の単語として収録されているものがある．例えば，「847　ひ　H= “火，灯” 」である．「日中辞典」によると，訳語が全く違っている．この場合，別々の単語とみなして扱うことにした．

2.2 中国語訳語の多義の構造

まず，双方の辞典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」を作成する．判別しやすいように，単語ごとに解釈語と例文を日本語で示す．この結果，60.1%の基本名詞に多義があり，最大多義数が 30，平均多義数が 2.6 であった．また，11 語が「日中辞典」では未収録である(図1)．

図1. 中国語訳語の多義の分布



2.3 日英の結果と比較

日本語及び英語について，5)によると IPAL 辞書における基本名詞の多義数及びその英語訳語の多義数(ALT-J/E における日英辞書)が報告されている．これと中国語の場合について表1に比較して表す．

表1 IPAL基本名詞の多義数について日英中の比較

	IPAL辞書	ALT辞書	日中辞典
平均多義数	2.13	1.88	2.6
最大多義数	18	12	30

3 意味属性の語義識別能力

3.1 意味属性による訳語選択

「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」[1]である．単語の「意味的用法」が単語の語義から派生することを考えると，実際に使われた文中での単語の「意味的用法」が分かれば，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に立つと期待できる．例えば，名詞「木」は，“树”，“木头”,“梆子” などという 3 つの中国語訳語があり，意味的用法としてそれぞれ「樹木」,「材木」,「楽器」という意味属性を持つ．そこで「木」の意味属性が決まることで，対応する訳語も決定できると考えられる．つまり,「木」の意味属性が「樹木」と決まれば“树”が,「材木」と決まれば“木头”が，そして「楽器」と決まれば“梆子”が選択できる．一方,「進歩」という日本語は中国語で“进步”,“长进”である．日本語には「進歩」という 1 つの意味属性しか定義されていないが，中国語では意味的用法が異なる．このように日本語では同義として用いられるため，意味属性による訳し分けはできない．

3.2 意味属性の適用性と語義識別能力の検討

3.2.1 意味属性の対応付け

まず，日中機械翻訳において,「日本語語彙大系」の「意味属性体系」がどの程度，適用するかを明らかするため，2.2 節の「対応表」の日本語の見出し語に対して，対応する中国語単語の語義の意味属性を調べて,「対応表」の「語彙大系の意味属性」という欄に追加記入した．

3.2.2 意味属性の量的な有効性

日本語名詞の意味属性の決まることで，中国語訳語の多義の解消の効果について調査した．日本語の見出し語 1,081 語に対して,その内訳を調べ，表 2 にまとめる．まず，訳語に意味属性が対応できたものについては，全体の 59.7%であった．49.6%は，意味属性と訳語の語義が 1 対 1 に対応し完全に多義が解消できる．10.0%は，訳語の語義全てに意味属性が対応付けられたものの，訳語が 1 つに絞り込めない．一方，意味属性の対応しない訳語がある場合は，全体の 40.1%である．意味属性が対応つけられない訳語は，訳出できないことになり，深刻な問題である．

表2 意味属性の語義識別能力

意味属性の対応しない訳語があるか	対応状況	件数	割合
全ての訳語に意味属性が対応	1対1で対応	531	49.6%
	重複し対応	108	10.0%
意味属性の対応しない訳語がある	意味属性が1つ不足	172	16.0%
	意味属性が2つ不足	100	9.3%
	意味属性が3つ不足	70	5.6%
	4つ以上不足	99	9.9%

3.2.3 訳語の多義から見た名詞の分類

次に,「対応表」から 2 つ以上の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．なお，638 語が対象となった．

そこで，日英翻訳における名詞分類方法[5]に習い，中国語訳語と意味属性との関係を調べたところ，表 3 に示すように分類された．同時にそれぞれの例も示す．なお，5)も同一の基本名詞を対象にしている．

表3 単語意味属性による名詞の分類

番号	分類	日英割合	日中割合	名詞の例		
				見出し語	意味属性	訳語
1	訳し分け可能	55%	15.5%	傘	「帽子」	斗笠
					「雨具」	伞
2	訳し分け一部可能		48.0%	顎	「あご」	(上下)腭
					「あご（頭部）」	下巴
						钓鱼钩的倒须
3	絞り込み可能	24%	5.3%	麻	「作物」	麻
					「作物」	大麻
					「繊維」	麻纤维
					「糸」	麻纱
					「布」	夏布
4	絞り込み不可能		11.1%	豆	「穀物」	豆
					「穀物」	黄豆
						腰子
5	訳し分け不可能	13%	18.8%	蜜柑	「果樹」「果物」	橘子
					「果樹」「果物」	柑橘
6	訳出不可能		0.5%	主	「正」	
						表面
						脸
7	未定義		0.6%	一言		一句话

*割合について空白の部分は，対応する分類観点がない．

4 日中翻訳のための意味属性体系

4.1 意味属性体系の識別能力からの考察

3.2節の結果から，訳語選択のできない原因を考察する．

原因 1. 日本語より中国語の語義が広い場合には，日本語名詞に元々つけられた意味属性では不足してしまう．

例えば，単語「手」は，日本語の場合には語義が 10，英語では 5[5]，そして中国語では 30 である．しかし，「意味属性体系」には，意味的用法として「594 手」，「1035 方法」，「1166 権利」「592 腕」という 4 つの意味属性しかない．このため日中翻訳の場合では，訳し分け能力が低い．

原因 2. 日本語より中国語の語義が具体的である場合には，訳出のために，適切な抽象度の意味属性を与えることが難しい．

例えば，日本語見出し語「鳥」に対応する意味属性は「538 鳥」であり，中国語訳語は「鸟」，「鸡」，「禽」などである．訳語選択時に意味的用法が指定されなければ，また，各訳語に具体的な意味属性が定義されていなければ，その訳語が区別できない．

原因 3. 訳語は名詞ではなく他の品詞であった．

例：見出し語「汗」の訳語の 1 つは「返潮」である．この意味は「湿る」などという意味となる．このときは名詞ではなく，動詞になった．

また，例文，仕事の傍ら勉強する．
そこで，単語「傍ら」の訳語は「一边…一边」である，これは副詞である．

原因 4. 本調査で 3 件程度であるが，訳語に対応する意味属性が全く存在しないものがあった．例えば，見出し語「おも」は，「表面」と「面孔」という訳語がある．「2439 正」という意味属性が定義されているが，訳語語義と全くあわない．

原因 5. 日中の文化的な差により，意味の包含関係が異なる．例えば，中国語では「饅頭類」の下位概念に「パン」がある．

4.2 意味属性体系の拡張

以上の原因を考慮して意味属性体系を拡張し，「対応表」の「拡張した意味属性」という欄に追加記入した．以下の通りに大別した：

（1）既存の意味属性の再配置

原因 1 の場合，「意味属性体系」から意味的用法に合う意味属性を日本語名詞に追加して割り当てる．同時にこの意味属性の子孫集合に，この単語を追加する．例えば，原因 1 の日本語の見出し語「手」を見てみる．意味属性は以下のように定義されている．

手(て)[名] 594 手 1035 方法 1166 権利 592 腕

中国語訳語の中では，「胳膊」や「（手里的）棋子」と「笔迹」などという意味がある．訳語の意味を考えると，意味属性「腕」で「胳膊」が決定できったのに対して，「（手里的）棋子」と「笔迹」は対応する意味属性がない．しかし，「意味属性体系」の全体を見ると，約 3000 個の意味属性のなかには，「921 遊び道具・運動具」と「1097 筆跡」という意味属性が定義されておく．そこで，見出し語「手」対して，2 つの意味属性を加え，それぞれ「遊び道具・運動具」と「筆跡」という「意味属性別単語表」に，単語「手」を同時に追加した．こうして，「意味属性体系」の意味素の数を変えることなく，日中翻訳の場合の識別能力が高められた．しかしながら，用言の結合価パターンの照合への影響が懸念される．

こうして，完全に多義が解消できた件数は 252 件だった．

（2）概括しすぎる意味属性の分解

表 3 の#4 など意味属性の細分化が足りない場合，訳語の語義を詳細に分析し，幾つかに分けてそれぞれに定義する．例えば，見出し語「鳥」を見れば，意味属性には，

「鳥[名] 538 鳥 843 肉・卵」

で定義されている．

一方，その訳語は「鸟」と「鸡」である．前者は「鸟」を「一般の鳥」とみなしている際の訳語であり，後者は「家畜鳥としての鳥」とみなしている際の訳語である．そこで，意味属性「鳥」から一部を抜き出し新しい意味属性「家禽類」に移す．そして，「鳥」の下位に「家禽類」を配置する（特に，ニワトリ・アヒルなどがこの類に属する）．

こうして，細分化した意味属性の数は 48 語であった．

（3）新しい意味属性の追加

具体的な例を 2 つ示す．例えば，日本語の見出し語「腹」の中国語訳語には，「肚儿」がある．これは

「指の腹」を表すが,対応する意味属性が存在しない．そこで以下の下線部分の通りに追加する．

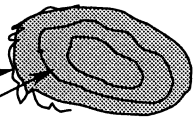
590 手・足
　591 手(上肢)
　599 足(下肢)
　608 指
　609 関節
　609zy 指のふくらみ

} この部分は，既存の「意味属性体系」である．

もう 1 つの例をあげる．意味属性「植物(部分)」には，「木質部(樹皮よりも木の中心部)」という意味属性を追加した．

686 植物(部分)
　687 芽・苗
　690 根
　691 茎・株
　694 枝・葉
　697 花
　700 実・種子・穂
　704 樹皮・果皮
　704zy 木質部（追加した意味属性）
　705 細胞

図 2．木の切断面



意味属性の増加と意味辞書の利用効率の関係調査については課題として残される．

こうして,新しく定義した意味属性の数は 10 語であった．

4.3 意味属性を拡張した効果

4.3.1 量的な有効性について

表 2 と同じ観点から，多義の解消の効果を表 3 に示す．一対一対応について，21.8%向上した．

表 4 拡張した意味属性の語義識別能力

意味属性が対応しない訳語があるか	対応状況	件数	割合
全ての訳語に意味属性が対応	1 対 1 で対応	773	71.5%
	重複し対応	91	8.4%
意味属性の対応しない訳語がある	意味属性が 1 つ不足	118	10.9%
	意味属性が 2 つ不足	53	4.9%
	意味属性が 3 つ不足	25	2.3%
	4 つ以上不足	21	1.9%

4.3.2 訳語の多義から見た名詞の分類

表 3 と同じ観点から，拡張した意味属性による訳語選択の可能性を検討した(表 5)．表 5 の#1 が高いことが最も望ましい．今回の意味属性の拡張により，15.5%から 59.5%に改善できた．この値は日英の場合より 4%も高い．

表 5 拡張した意味属性による名詞の分類

番号	分類	日英割合	日中割合	
			前	後
1	訳し分け可能	55%	15.5%	59.5%
2	訳し分け一部可能		48.0%	2.7%
3	絞り込み可能	24%	5.3%	22.8%
4	訳し分け不可能	13%	18.8%	14.7%
5	未定義		0.6%	0.5%

5 おわりに

本稿では，日中機械翻訳において，意味属性を用いた日本語の基本名詞の訳語選択の可能性を検討した．IPAL 辞書の 1,081 語の基本名詞を対象に選択能力を検討したところ．49.6%の見出し語は，現在の意味属性体系により完全に訳語が選択できることが分かった．訳語選択に失敗した例を分析し，「意味属性体系」を拡張した．その結果，71.5%の見出し語が完全に訳語選択できるように改良された．これは，日英における訳語選択能力よりの高い値である．意味属性を拡張する際，日中の文化的差により構造を変更する必要はあったが，この値より，日英よるも日中の方が近い語義のあることを表しているのではないかと考える．

今後の課題は，訳語の語義の抽象度が高い場合，意味属性によって区別できなかったことを検討することである．

参考文献

1)池原，他(1997)：日本語語彙大系　1．意味体系，岩波書店．

2)金出地，徳久，池原，村上(2003)：結合価文法による動詞と名詞の訳語選択能力の評価，自然言語処理研究会，2003-NL-153-16,pp119-124.

3)計算機用日本語名詞辞書 IPAL 解説編(1996)，情報処理振興事業協会技術センター．

4)**宋文军**，他(1987)，北京・対外経済貿易大学と北京・商務印書館及び小学館の共同編集．

5)桐沢，池原，村上(2000)：日英機械翻訳における名詞の訳語選択，言語処理学会，B1 - 6，pp.55 - 88