日本語機能表現の用法識別および英訳規則獲得

阿部佑亮　(筑波大学大学院システム情報工学研究科)
鈴木敬文　(筑波大学大学院システム情報工学研究科)
宇津呂武仁　(筑波大学大学院システム情報工学研究科)*
松吉俊　(山梨大学大学院医学工学総合研究部)

Discriminating Usages of Japanese Functional Expressions and Acquiring Rules for Translation into English

Yusuke Abe　(University of Tsukuba)
Takafumi Suzuki　(University of Tsukuba)
Takehito Utsuro　(University of Tsukuba)
Suguru Matsuyoshi　(University of Yamanashi)

1. はじめに

機能表現とは，以下の例文の「について」,「にちがいない」,「とはいえ」のように複数の語が 1 つの助詞・助動詞・接続詞のようにふるまう表現を指す．機能表現は，表現全体で 1 つの非構成的意味を持つという特性を持つ.

- 格助詞型 農村の生活 **について** 調べている.
- 助動詞型 これは天狗の仕業 **にちがいない**.
- 接続詞型 手紙を出した **とはいえ**，返事が来るとは限らない.

本研究の目的は，これら機能表現を正しく翻訳する仕組みを提案することである．しかし，日本語機能表現には，多様な異形が多く存在するが，現状の日英機械翻訳ソフトにおいて，それらの異形を網羅的に正しく翻訳することは容易ではない [坂本 09]．[山本 01] では，原言語における類似の表現を，代表的な表現に言い換えた後，機械翻訳の言語変換部を適用するという SandGlass 翻訳方式を提案している．本研究では，この SandGlass 翻訳方式の考え方をふまえて，機能表現を意味ごとにまとめ，その意味のまとまりを単位として翻訳規則の獲得を行う.

本研究で対象とする日本語機能表現の情報として，日本語機能表現を網羅的に列挙した大規模日本語機能表現階層辞書「つつじ」 [松吉 07,松吉 08](16,801 表現を収録)，および,「つつじ」に収録されている機能表現を言い換え可能な表現ごとに階層的に分類した「意味的等価クラス」 [松吉 08](最下層には全 199 の意味的等価クラスが存在) を利用する．前述の意味のまとまりの単位として，意味的等価クラスを用いる．また，翻訳規則を獲得するために，NTCIR-7 の特許翻訳タスク [Fujii08] で配布された 1,798,571 件の日英対訳特許文対を用いる．これらに句に基づく統計的機械翻訳システム Moses [Koehn03] を適用し，学習されたフレーズテーブルから翻訳規則を獲得する．特許文の場合は，使用される機能表現の意味範囲が狭く，その種類も少ないので，翻訳規則の獲得が容易である点が大きな利点となる.

ここで，特許文を情報源として英訳規則を獲得するという課題において，本研究では，まず，1 つの意味的等価クラスにおいて，日本語機能表現の複数の用法を識別するという点について考える．また，それに加えて,「つつじ」に収録されている各意味的等価クラスに対して，意味のまとまりを考

*utsuro @ iit.tsukuba.ac.jp

表 1: 特許文を対象とした日本語機能表現の英訳規則獲得：用法識別の必要性・意味的等価クラス単位での集約の効果

		意味的等価クラス単位での集約の効果	
		効果が大きい	効果が小さい
日本語機能表現の用法識別の必要性	必要	分類 (a) (本稿で対象とする 4 クラスを含む)	分類 (b) (本稿で対象とする 5 クラスを含む)
	不要	分類 (c)	分類 (d)

慮した集約的英訳の考え方がどの程度適用できるかについても考える．この 2 つの観点から意味的等価クラスを，4 通り (各観点 2 通りの分類× 2 観点) に分類をしたものを，表 1 に示す．

まず，1 つの意味的等価クラスにおける日本語機能表現の用法の識別としては，大きく分けて 2 種類のものがある．1 つは，文中の表現が機能表現の意味として用いられているもの (機能的用法) と，その表現を構成する語本来の意味で用いられているもの (内容的用法) との間の用法識別，もう 1 つは，機能表現の意味が文脈によって異なるという機能的用法の用法識別である．より正確な英訳を行うためには，これらの機能表現の用法の識別を考慮した英訳規則の獲得の仕組みが必要である．本稿では，用法識別が必要な日本語機能表現のうちの 9 クラス (表 1 の分類 (a) に含まれる 4 クラスと，分類 (b) に含まれる 5 クラス) を対象とした．

次に,「つつじ」に収録されている各意味的等価クラスに対して，意味のまとまりを考慮した集約的英訳の考え方がどの程度適用できるかについて考えるために，1 つの意味的等価クラスあたり何種類の日本語機能表現が特許文において出現するかを調べた．NTCIR-7 の特許翻訳タスクで配布された日英対訳特許文対において，各意味的等価クラスに属する機能表現の出現種類数を調査したところ，全 199 個の意味的等価クラスのうち，1 クラスあたり 10 種類以上の機能表現が出現している意味的等価クラスは，38 クラスであった．その他の 161 個の意味的等価クラスについては，集約の効果が小さいと考えられる．これらの 161 クラスは，表 1 の分類 (b)，分類 (d) のいずれかに分類される．一方，1 クラスあたり 10 種類以上の機能表現が出現し，集約の効果が大きい 38 クラスは，表 1 の分類 (a)，分類 (c) のいずれかに分類される[1] ．

本稿では，用法識別が必要な意味的等価クラスのうち，9 クラスを対象として用法識別を行い，それにもとづく英訳規則の獲得を行う．用法識別に関して，[鈴木 11] では，文中に出現した日本語機能表現の表記について，機能的用法と内容的用法とを，当該表記およびその周辺の語の形態素情報を用いて判別する手法を提案している．本稿では， [鈴木 11] の手法を機能的用法・内容的用法の用法識別と機能的用法の用法識別の両方に対して適用する．

2. 階層的日本語機能表現辞書

[松吉 07] は，日本語機能表現を各表現の構成要素の組み合わせとして階層的に網羅した辞書を作成した (日本語機能表現一覧「つつじ」[2]). この辞書は [土屋 06] の用例データベースを受けて，辞書に収録する機能表現の範囲を拡張することを目指したもので，日本語の機能表現の異型を，機能表現の構成要素の組み合わせとして階層的に収録している．これにより，日本語機能表現の網羅的取り扱いが可能となった．この辞書には，機能表現末尾の活用だけでなく，機能表現の各構成要素の音韻的変化や，とりたて詞の挿入，口語的な表現と敬語表現の差し替えなどによる異型を機械的に展開した後に，実際に日本語として使用できるものだけを人手で残した 16,801 表現が収録されている．

[1]集約をすることの意義は，まず，英訳規則を減らすことにある．そのため，機能表現の種類数が少ないと集約の効果が小さい．また，本研究では，句に基づく統計的機械翻訳システムによって学習されたフレーズテーブルから獲得された英訳規則を用いているが，一般的な傾向として，統計的機械翻訳の翻訳精度は，低頻度の表現に対しては低いことが予想される．そこで，意味のまとまりを考慮した集約的英訳によって，低頻度の表現の英訳にも対応でき，精度改善が見込める，という点が集約のもう一つの意義として挙げられる．

[2]http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

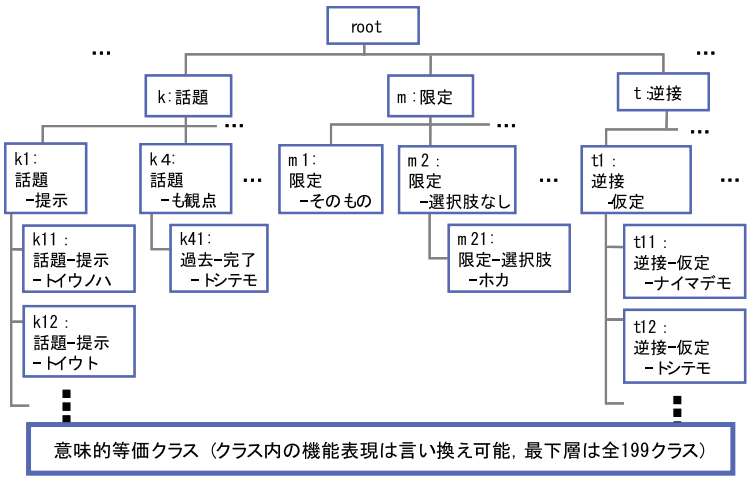

図 1: 「つつじ」の意味的等価クラス

表 2: 日本語機能表現の多義性の例

(a) 多義性を持たない日本語機能表現

	表現	例文	用法
(a1)	ことができる	また設計上の必要に応じて第１および第２の部材２２，２３の寸法，形状を変えれば，設計上の適正値を実現することができる．	機能的用法 E11(可能-可能-コトガデキル)

(b) 機能的用法・内容的用法の多義性を持つ日本語機能表現

	表現	例文	用法
(b1)	ものの	乾燥に供した加熱空気は蒸発した水蒸気を含み，多くの熱エネルギーを持っているものの，回収して循環利用するには限界があり，多くの場合廃棄されている．	機能的用法 t24(逆接-確定-モノノ)
(b2)	ものの	ここで，ブロックが存在しない場合は，探索対象段の位置を，保持されたアベイラブルエリアで最後のものの左上隅点とし（ステップ１１０６），その後，後述する図１２に示される処理を実行する。	内容的用法

(c) 機能的用法の多義性を持つ日本語機能表現

	表現	例文	用法
(c1)	としても	このため、誤って装置に物等を落下したとしても，その衝撃は反射ミラー８ｆに伝わり難くなっている．	機能的用法 t12(逆接-仮定-トシテモ)
(c2)	としても	さらに，ブレード４５は接触ローラ３７の外周面３７ａの汚れを除去するクリーニング手段としても作用する．	機能的用法, k41(話題-も観点-トシテモ)

また，[松吉 08] は，上記の辞書に収録された見出し語間の類似度に応じて，3 段階のクラス分けをおこなった．図 1 に示すように，上記の辞書に収録されている見出し語は階層的に意味的等価クラスに割り振られている．この最下層に位置する全 199 個の各意味的等価クラスに属する機能表現群は，日本語文中で言い換え可能であるとされている．

3. 日本語機能表現の多義性

1 節でも述べたように，本研究の目的は，対訳特許文を利用した日本語機能表現の集約的英訳規則の作成である．しかし，集約的英訳規則を作成するにあたって，日本語機能表現の持つ多義性を解消する必要がある．

日本語機能表現には，大きく分けて 2 種類の多義性がある．1 つは，文中の表現が機能表現の意味

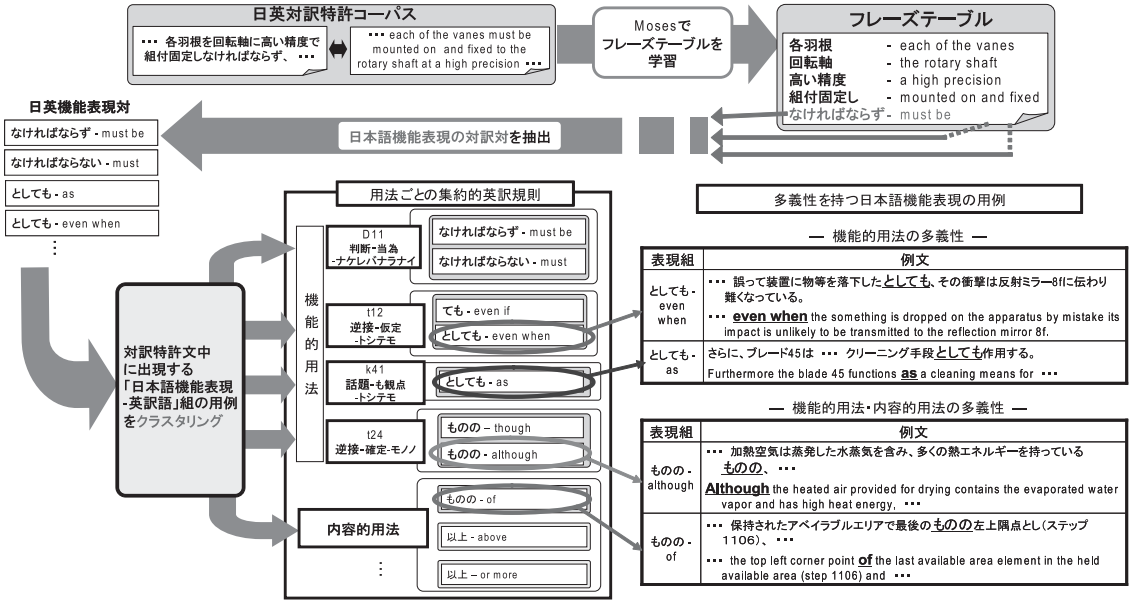

図 2: 日本語機能表現の用法識別および用法識別を考慮した集約的英訳規則作成

で用いられる場合 (機能的用法) と，その表現を構成する語本来の意味で用いられる場合 (内容的用法) の間での多義性である．もう 1 つは，機能表現の意味が文脈によって異なるという機能的用法の多義性である．これらの多義性の例を，表 2 に示す．(a) は，多義性を持たない機能表現の例である．一方，(b)，(c) は多義性を持つ機能表現の例であり，(b) は機能的用法 (b1) と内容的用法 (b2) の例を，(c) は機能表現として異なる意味で用いられている例を，それぞれ示している．

4. 日本語機能表現の用法識別および集約的英訳規則作成

日本語機能表現の用法識別，および，用法識別を考慮した集約的英訳規則作成の様子を図 2 に示す．

まず，[島内 11] と同様にして，句の日英対応およびその確率を記載したフレーズテーブル[3]を作成し，そこから，大規模日本語機能表現階層辞書「つつじ」[松吉 07] に収録されている機能表現のエントリを含む日英機能表現対の用例を抽出する[4]．

次に，抽出された日本語機能表現の表記が出現する対訳特許文の集合を用いて,「日本語機能表現-英訳語」組の用例のクラスタリングを行う．これによって，対訳特許文中に出現している各「日本語機能表現表記-英訳語」組を用法・意味ごとに分類すると同時に，翻訳規則として集約することができる．図 2 では，機能表現表記「ものの」について,「ものの-though/although」は機能的用法として「t24(逆接-確定-モノノ)」に,「ものの-of」は内容的用法として分類され，機能的用法の”though” と”although” は同様の意味なので，1 つの翻訳規則として集約されている．また，機能表現表記「としても」について,「としても-even when」は「逆接仮定」の意味なので「t12(逆接-仮定-トシテモ)」に,「としても-as」は「観点」の意味なので「k41(話題-も観点-トシテモ)」に分類され,「t12(逆接-仮定-トシテモ)」では「ても-even if」と同様の意味なので，1 つの翻訳規則として集約されている．

本稿においては，日英機能表現対の用例のクラスタリングにおいて，日本語用例，および，その前後の文脈間の類似度を定義し，各用例における日本語機能表現の用法が類似する用例のクラスタリングを行った．類似度の算出においては，日本語機能表現表記を構成する形態素，および，その出現位置の前後の形態素の品詞およびその活用形を利用した [鈴木 11]．具体的には，用例中の日本語機

[3]句に基づく統計的機械翻訳モデル [Koehn03] のツールキットである Moses を適用することにより作成した．

[4]対訳特許文対における日本語機能表現の出現頻度の下限を 20，対訳特許文対における日本語機能表現および英訳語が句対応していると判定された頻度の下限を 10，フレーズテーブルにおける日英翻訳確率 $P(f_e \mid f_j)$(日本語フレーズ f_j が英語フレーズ f_e に翻訳される条件付確率の形式) の下限を 0.05 とする．

表 3: 日本語機能表現の用例間の類似度算出の例

	機能表現表記	例文	用法	形態素情報		
				前接形態素	機能表現表記を構成する形態素	後接形態素
例 1-1	ことができ	このようにウェビングを形成することにより、エアベルトの同じ効果を得る[ことができ]ながら、ガスの発生量を小さくすることができる。	機能的用法 E11 (可能-可能-コトガデキル)	得る: 動詞/自立/基本形	こと:名詞/非自立/＊ ＋ が:助詞/格助詞/＊ ＋ でき:動詞/自立/連用形	ながら: 助詞/接続助詞/＊
例 1-2	ことができれ	またチップ番号 z によって、ロット番号，ウエハ枚数，ウエハ番号等を表示する[ことができれ]ば、それらの情報はなくてもよい。	機能的用法 E11 (可能-可能-コトガデキル)	する: 動詞/自立/基本形	こと:名詞/非自立/＊ ＋ が:助詞/格助詞/＊ ＋ できれ:動詞/自立/仮定形	ば: 助詞/接続助詞/＊
例 1-1，および，例 1-2 間の類似度		各項目間の個別類似度		3	(3+3+2)/3 ≈ 2.7	3
		用例全体の類似度		8.7		
例 2-1	としても	従ってこれらの部品の形状付与のためには、成形品の特性は劣る[としても]、例えば鋳造などの別の方法を採らざるを得ない場合があった。	機能的用法 t12 (逆接-仮定-トシテモ)	劣る: 動詞/自立/基本形	として:助詞/格助詞/＊ ＋ も:助詞/係助詞/＊	、: 記号/読点/＊
例 2-2	としても	また、収納 y アーム 5 1 についても、吸着部を n 個単位で設ける構成[としても]良く、同様にして実施でき、より高速のサイクルタイムに対応可能になる。	機能的用法 k41 (話題-も観点-トシテモ)	構成: 名詞/サ変接続/＊	として:助詞/格助詞/＊ ＋ も:助詞/係助詞/＊	良く: 形容詞/自立/連用テ接続
例 2-1，および，例 2-2 間の類似度		各項目間の個別類似度		0	(3+3)/2 = 3.0	0
		用例全体の類似度		3.0		

能表現表記の構成形態素列を M_c, 前接する形態素を m_{pre} , 後接する形態素を m_{suf} として，用例 e を $\langle m_{pre}, M_c, m_{suf} \rangle$ で表す．そして，2 つの用例 e_1, e_2 の間の類似度 $Sim(e_1, e_2)$ を,「日本語機能表現表記を構成する形態素とその前後の形態素それぞれの品詞，品詞細分類，活用形の一致数の合計」として定義した[5]（類似度の最大値は 9 となる）.

$$Sim(e_1, e_2) \quad = \quad Sim_{pre}(m_{pre}(e_1), m_{pre}(e_2)) \ + \ Sim_c(M_c(e_1), M_c(e_2)) \ + \ Sim_{suf}(m_{suf}(e_1), m_{suf}(e_2))$$

日本語機能表現の用例間の類似度算出の例を表 3 に示す．例 1-1，例 1-2 は，どちらも「可能」の用法として用いられた機能表現の用例である．機能表現表記は異なるが，その表記を構成する形態素やその前後の形態素の品詞及び活用形が類似しているため，類似度は 8.7 という大きい値となる．一方，例 2-1，例 2-2 は，機能表現表記は同一であるが異なる用法 (例 2-1 は「逆接仮定」，例 2-2 は「観点」) で用いられている用例である．前後の形態素の品詞およびその活用形が全く異なるため，類似度は 3 という小さい値となる．以上の類似度に基づいて，ボトムアップクラスタリングを行うことで，より効率的に集約的英訳規則を作成することが可能となる．

5. 評価

本節では，4 節で述べた手法を用いてクラスタリングを行った結果を評価する．その際,「クラスタリングによる日本語機能表現の用法識別結果」と「集約的英訳規則作成」という観点で評価・分析を行った．ただし，本稿では [島内 11] で英訳規則作成時に用いた 53 の意味的等価クラスの内，用法の分布が偏っている 30 クラスを除いた 23 クラスを選出した．そのうち，用法が適度に混在しており，意味的等価クラス単位での集約の効果が大きい 4 クラスと，効果が小さい 5 クラスの，計 9 クラスに属する機能表現を含む用例を評価の対象とした[6].

[5]日本語機能表現表記を構成する形態素数が 2 形態素以上の場合，構成形態素の数で一致数を正規化している．

[6] [島内 11] では，本研究とは逆に，当該クラスの用法の用例が偏って出現する 30 クラスの内，20 クラスを対象として集約的英訳規則の作成を行っている．

表 4: クラスタリングによる用法識別，および，集約的英訳規則作成の評価結果

意味的等価クラス		用例数	クラスタ数	精度 (%)
意味的等価クラス単位での集約の効果が大きい	J33(進行-継続-テクル)	25	9	92.1
	J31(進行-継続-テイル)	61	8	90.8
	M11(不必要-不必要-ナクテヨイ)	66	5	77.9
	t12(逆接-仮定-トシテモ)	37	6	70.7
意味的等価クラス単位での集約の効果が小さい	n12(添加-非限定-ダケデナク)	31	3	100
	P11(例示-程度-クライ)	12	5	100
	c11(仲介-原因-ニヨッテ)	56	8	77.5
	s11(理由-因状況-イジョウハ)	15	4	70.0
	m12(限定-そのもの-ノミ)	20	6	65.6

まず，クラスタリングによる日本語機能表現の用法識別，および，集約的英訳規則作成の評価結果を表 4 に示す．表中の精度は,「生成された各クラスタ内の用例において『日本機能表現 - 対訳英語』対が同じ用法である割合」のミクロ平均を示している．また，クラスタ数ついては，日本語機能表現の多義性，および，英語訳語の用法の違いを適切に分類できているかという観点から人手で決定した.「n12」,「P11」,「J33」,「J31」の 4 クラスに関しては，4 節で述べた類似度尺度が有効に働き，適切な分類ができていた．一方,「M11」,「c11」,「t12」,「s11」,「m12」の 5 クラスに関しては，用法の異なる用例間で類似度が大きくなり，十分な分類ができなかった．改善策としては，日本語の係り受け関係や日本語機能表現に対応する英語表記などを利用することが考えられる．

6. まとめと今後の課題

本稿では，対訳特許文を対象として，日本語機能表現の用例をクラスタリングすることにより, 日本語機能表現の用法を識別し，集約的英訳規則を作成する手法を提案した．今後の課題として，まず，英語訳語の用法の類似性を考慮したクラスタリングを行うことにより，集約的英訳規則を作成する．また，各機能表現の全用例を対象としたクラスタリングを行う．

参考文献

[Fujii08] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the Patent Translation Task at the NTCIR-7 Workshop, *Proc. 7th NTCIR Workshop Meeting*, pp. 389–400 (2008).

[Koehn03] Koehn, P., Och, F. J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. HLT-NAACL*, pp. 127–133 (2003).

[松吉 07] 松吉俊, 佐藤理史, 宇津呂武仁：日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123–146 (2007).

[松吉 08] 松吉俊, 佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え, 自然言語処理, Vol. 15, No. 2, pp. 75–99 (2008).

[坂本 09] 坂本明子, 宇津呂武仁, 松吉俊：日本語機能表現の集約的英訳, 言語処理学会第 15 回年次大会論文集, pp. 654–657 (2009).

[島内 11] 島内蘭, 阿部佑亮, 鈴木敬文, 宇津呂武仁, 松吉俊：特許文における日本語機能表現の集約的英訳規則の作成と評価, 言語処理学会第 17 回年次大会論文集, pp. 396–399 (2011).

[鈴木 11] 鈴木敬文, 宇津呂武仁, 松吉俊, 土屋雅稔：代表・派生関係および用例を利用した日本語機能表現の解析, 言語処理学会第 17 回年次大会論文集, pp. 155–158 (2011).

[土屋 06] 土屋雅稔, 宇津呂武仁, 松吉俊, 佐藤理史, 中川聖一：日本語複合辞用例データベースの作成と分析, 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741 (2006).

[山本 01] 山本和英, 白井諭, 坂本仁, 張玉潔：SANDGLASS: 両言語換言機構を基軸とする音声翻訳, 言語処理学会第 7 回年次大会発表論文集, pp. 221–224 (2001).