対訳用例および意味的等価クラスを用いた機能表現の日英翻訳*

阿部 佑亮† 鈴木 敬文† 宇津呂 武仁‡ 山本 幹雄‡ 松吉 俊§ 河田 容英¶
筑波大学大学院 システム情報工学研究科† 筑波大学 システム情報系‡
山梨大学大学院 医学工学総合研究部§ (株) ナビックス¶

1 はじめに

日本語には 16,000 種類以上の機能表現 (助詞・助動詞・接続詞相当語句) の異形が存在する．日本語機能表現には非常に多様な異形が存在するが，それらの異形を網羅的に正しく翻訳することは難しい．この問題に対応する手法として，先行研究では，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における機能表現の意味的等価クラスを利用して，日英対訳特許文中に出現する日本語機能表現の日英翻訳を対象として，日本語機能表現の集約的な日英機械翻訳を行う手法を提案している．この手法を 53 の意味的等価クラスに適用した結果，20 クラスについては，意味的等価クラスに属する日本語機能表現の翻訳規則を 1 規則ないし 2 規則に集約出来ることが分かった．しかし，一方で，他の 33 クラスについては，意味的等価クラスに属する日本語機能表現の翻訳規則を集約することが出来なかった．これは，日本語機能表現を英訳する際の曖昧性のためであった．より正確な翻訳を行うためには，これら機能表現表記のもつ曖昧性を考慮した翻訳の仕組みが必要不可欠である．

以上を踏まえて，本論文では，NTCIR-7 の特許翻訳タスクで配布された 1,798,571 件の日英対訳特許文対から得た対訳用例を用いて，日本語機能表現を英訳する方式を提案する．この方式においては，機能表現の意味的等価クラスごとに，様々な対訳用例からデータベースを構築し，英訳対象となる機能表現表記の用例と最も類似した対訳用例の訳語を適用することで，上記の曖昧性に対応する．評価実験として，句に基づく統計的機械翻訳モデル Moses [3] を，日英対訳特許文を用いて訓練したものとの翻訳精度比較を行った．両手法の作成時に参照するテキストと同ジャンルである特許文における翻訳精度は，多くの意味的等価クラスにおいて Moses の方が優れていたが，「日本語書き言葉均衡コーパス」および「日本語学習者用用例集」における翻訳精度は，多くの意味的等価クラスにおいて提案手法の方が優れていた．このことから，対訳用例を選定したテキストとは異なるジャンルのテキストにおける英訳においても，提案手法は比較的安定した翻訳性能を示すことを実証できた．

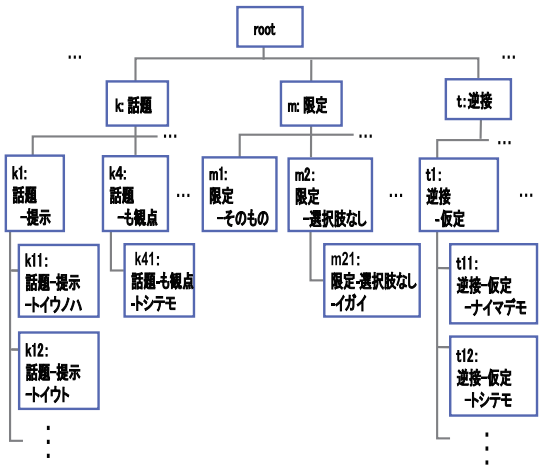

図 1: 意味的等価クラスに基づく階層構造

2 階層的日本語機能表現辞書

文献 [6, 5] は，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た辞書を編纂した (日本語機能表現一覧「つつじ」[1])．日本語機能表現一覧「つつじ」には，16,801 の機能表現が収録されており，この辞書によって，日本語機能表現の網羅的取り扱いが可能になった．

また，日本語機能表現一覧「つつじ」では，図 1 に示すように，辞書に収録されている見出し語について，意味的等価クラスという形での階層的分類も行っている．この最下層に位置する全 199 個の意味的等価クラスについて，同一クラス内の機能表現は，日本語文中で言い換え可能であるとされている [5]．この意味的

*Japanese to English Machine Translation of Functional Expressions based on Translation Examples and Semantic Equivalence Classes

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

表 1: 機能表現表記の曖昧性の例

(a) 機能的用法/内容的用法の曖昧性

	表記	例文	用法
(1)	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) (～ ものの= *although* ～)
(2)	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 (～ ものの = *of*)

(b) 複数の機能的用法間の曖昧性

	表記	例文 (英訳文)	用法
(3)	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) (としても= *even when*)
(4)	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) (としても= *as*)

(c) 対訳英語の曖昧性

	表記	例文 (英訳文)	用法
(5)	による	原稿台１１側からの光のミラー１４ による 反射光路上には結像レンズ１６とプラテン２０がこの順に配置されている。	機能的用法 c11(仲介-原因-ニヨッテ) (による = *by*)
(6)	による	本発明 による 可変差動制限装置２の制御は、以下の (1)，(2)，(3) の３種の制御の組合せから構成される。	機能的用法 c11(仲介-原因-ニヨッテ) (による= *according to*)
(7)	による	つまり、放電開始 による 電圧の低下が、極間異常状態と判定されてしまうことがある。	機能的用法 c11(仲介-原因-ニヨッテ) (による= *due to*)

等価クラスを用いることにより，日本語機能表現の言い換え候補を網羅的に取り扱うことが可能となった．

3 機能表現表記の曖昧性

日本語機能表現表記の適切な英訳を行うためには，日本語機能表現表記の持つ曖昧性に対応する必要がある．日本語機能表現表記を英訳するにあたって，対応すべき曖昧性は，大きく分けて 3 種類ある．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) との間の曖昧性である (表 1 (a))．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曖昧性である (表 1 (b))．そして最後の 1 つは，対訳英語の曖昧性である (表 1 (c))．

4 対訳用例データベースの構築

本論文では，NTCIR-8 の特許翻訳タスク [1] で配布された日英対訳特許文の文対応データのうち，1,798,571 件をフレーズテーブルの訓練用データとして使用した．この文対応データに対して，句に基づく統計的機械翻訳モデルのツールキットである Moses [3] を適用し，日英の句の組および日英の句の組が対応する確率を示したフレーズテーブルを作成する．

このフレーズテーブルを用いて，先の約 180 万件の日英対訳特許文対から，対訳用例データベースを構築した．具体的には，対訳用例データベースの構築対象としている意味的等価クラスに属する各日本語機能表現表記について，以下の条件を満たす「日本語機能表現表記- 英訳語」組をフレーズテーブルから抽出する．

- 日英対訳特許文対における日本語機能表現表記の出現頻度が 20 以上
- Moses によって，日英対訳特許文対における「日本語機能表現表記 - 英訳語」組が句対応していると自動判定された箇所の頻度が 10 以上
- フレーズテーブルにおける日英翻訳確率が 0.05 以上

そして，抽出した各「日本語機能表現表記 - 英訳語」組について，この表記および英訳語が対応関係であると人手で判断された対訳文対を，約 180 万件の日英対訳特許文対から収集し，対訳用例データベースへ登録する．ただし，本研究では，意味的等価クラスごとのデータベースを用意し，当該日本語機能表現が属する意味的等価クラスのデータベースにのみ，対訳用例を追加する．この構築手順に従って，表 2 に示した 10 の

表 2: 対訳用例データベースを構築した意味的等価クラス

意味的等価クラス			表現数	表現の例
日本語機能表現表記の用法の曖昧性	大きい	M11(不必要 - 不必要 - ナクテヨイ)	299	なくてもよい，までもない，ずともよく
		P11(例示 - 程度 - クライ)	6	くらい，ばかり，ほど
		c11(仲介 - 原因 - ニヨッテ)	15	により，をもって，によります
		m12(限定 - そのもの - ノミ)	5	きり，だけ，のみ
		n12(添加 - 非限定 - ダケデナク)	12	のみならず，だけじゃなく，上に
		s11(理由 - 因状況 - イジョウハ)	9	からには，うえは，以上
		t12(逆接 - 仮定 - トシテモ)	21	にしても，としましても，たところで
	小さい	D11(判断 - 当為 - ナケレバナラナイ)	213	ないといけない，ねばならない，べき
		b11(対象 - 関連 - ニツイテ)	26	に関する，について，につきまして
		u12(対比 - 般 - カワリニ)	14	代わりに，代りに，かと思うと

意味的等価クラスの対訳用例データベース構築を行った結果，10 クラス合計で 5,253 用例の対訳用例データベースを構築することができた．

5 対訳用例を用いた機能表現の日英翻訳

入力された日本語用例を $e_j = \langle m_{pre}, M_c, m_{suf} \rangle$，その日本語用例 e_j 中の日本語機能表現表記を $f_j(e_j)$，データベース中のある用例を $e_{je}^{db} = \langle e_j^{db}, t_e^{db} \rangle$ とする．m_{pre}，m_{suf} はそれぞれ，日本語機能表現表記に前接する，あるいは，後接する形態素を表し，M_c は，日本語機能表現表記を構成する形態素列を表す．提案手法は，e_j と e_{je}^{db} との類似度 $Sim(e_j, e_{je}^{db}) = Sim(e_j, \langle e_j^{db}, t_e^{db} \rangle)$ が最大となる用例の英訳語 t_e^{db} を選択し，日本語用例 e_j 中の日本語機能表現表記 $f_j(e_j)$ に対する英訳語として出力する．ただし，類似度 $Sim(e_j, \langle e_j^{db}, t_e^{db} \rangle)$ は以下で定義される．

$$
\begin{aligned}
&Sim\left(e_j, e_{je}^{db} = \left\langle e_j^{db}, t_e^{db} \right\rangle\right) \\
&= Sim_{pre}\left(m_{pre}(e_j), m_{pre}(e_j^{db})\right) \\
&\quad + Sim_c\left(M_c(e_j), M_c(e_j^{db})\right) \\
&\quad + Sim_{suf}\left(m_{suf}(e_j), m_{suf}(e_j^{db})\right)
\end{aligned}
$$

ここで，Sim_{pre} および Sim_{suf} はそれぞれ，e_j，e_j^{db} の前接形態素，および，後接形態素の類似度であり，Sim_c は，e_j の構成形態素列と e_j^{db} の構成形態素列の類似度である．提案手法による訳語選択の例を，図 2 に示す．

6 評価

提案手法の評価として，提案手法および句に基づく統計的機械翻訳モデルである Moses [3] を日英対訳特許文を用いて訓練したものについて，翻訳精度の比較を行った．評価は，4 節で述べた手順に従って対訳用例データベースを構築した 10 の意味的等価クラスを対象として行った．

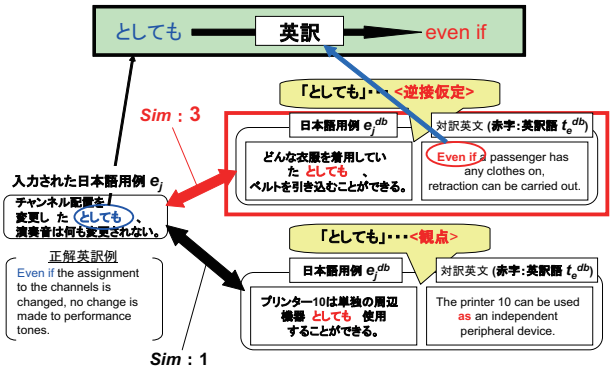

図 2: 日本語機能表現表記の用例間の類似度を用いた訳語選択

評価文は，NTCIR-8 の特許翻訳タスク [1] で配布された日英対訳特許文のうちの約 140 万件 (以下,「特許文」),「現代日本語書き言葉均衡コーパス」 [4](以下,「書き言葉コーパス」)，および,「日本語学習者用用例集」 [2]，の 3 種類のテキストから収集した．また，評価文は,「対訳用例データベース中の各用例との類似度の最大値 $Sim_{max}(e_j)$ [2]の範囲」および「評価文 e_j 中の日本語機能表現表記 $f_j(e_j)$ が Moses の学習データである約 180 万件の対訳特許文中に出現するか」の 2 つの尺度で分類した．前者に関しては，$2.33 \leq Sim_{max}(e_j) \leq 3$ を満たす評価文 e_j を「提案手法での英訳が容易」，$0 \leq Sim_{max}(e_j) < 2.33$ を満たす評価文 e_j を「提案手法での英訳が容易でない」と，それぞれ判断した．同様に後者に関しては，日本語機能表現表記 $f_j(e_j)$ が Moses の学習データである約 180 万件の対訳特許文中に出現する評価文 e_j を「Moses での英訳が容易」，出現しない評価文 e_j を「提案手法での英訳が容易でない」と，それぞれ判断した．この 2 つの尺度に従って，各意味的等価クラスについて 4 種類の評価文集合を作成した．また，評価文は，各評価文集合から選定された評価文の数が均等になるように選定した．

[2] $Sim_{max}(e_j) = \max_{e_{je}^{db}} Sim(e_j, e_{je}^{db} = \langle e_j^{db}, t_e^{db} \rangle)$ と定義する．

表 3: 評価結果

評価対象		評価文数	翻訳精度 (%)			
			ベースライン	Moses	提案手法	提案手法 (上限値)
テキストジャンル別	特許文	140	53	**66**	65	70
	書き言葉コーパス	292	53	34	**77**	78
	日本語学習者用用例集	168	51	42	**56**	62
評価文集合別	評価文集合 (1) (提案手法：英訳が容易，Moses：英訳が容易)	210	54	**66**	**66**	73
	評価文集合 (2) (提案手法：英訳が容易でない，Moses：英訳が容易)	180	48	47	**54**	58
	評価文集合 (3) (提案手法：英訳が容易，Moses：英訳が容易でない)	90	58	31	**74**	82
	評価文集合 (4) (提案手法：英訳が容易でない，Moses：英訳が容易でない)	120	55	20	**63**	65
合計		600	53	46	**63**	69

評価結果を表 3 に示す.「提案手法」,「Moses」は，それぞれの手法での翻訳精度を示しており,「提案手法 (上限値)」は，提案手法において複数の英訳語が出力され，その中に評価文中の機能表現表記に対して適切な英訳語と不適切な英訳語が混在する場合，その中から適切な英訳のみを選択し，出力した場合の翻訳精度を示している.「ベースライン」は，各意味的等価クラスの対訳用例データベースにおける対訳用例の英訳語 t_e^{db} のうち，頻度最大のものを英訳語として出力した場合の精度である．評価の結果，評価文書集合全体では，提案手法の方が Moses よりも優れた翻訳精度となった．また，ベースラインと比較して，どの評価文集合においても，提案手法の方が翻訳精度が高いことから，提案手法によって機能表現表記の曖昧性に対応した翻訳が行われていることが分かる．

テキストのジャンル別の翻訳精度を見てみると，両手法の作成時に参照するテキストと同ジャンルである「特許文」では，Moses の方が精度が高いが，それ以外のジャンルのテキストでは，提案手法の方が精度が高いことが分かる．このことから，提案手法は，特定のジャンルのテキストから対訳用例を集めても，それとは異なるジャンルのテキストへ適応できる可能性が高いことが分かる．通常，一般的なジャンルにおいては，二言語対訳コーパスの作成は高コストであり，使用できるものが限られるため，この特性は重要であると考えられる．

評価文中の日本語機能表現表記が訓練データ中に出現しない評価文集合 (3)，(4) における翻訳精度の比較結果をふまえると，Moses は訓練データから作成したフレーズテーブルを用いて翻訳を行うため，訓練データ中に出現しない表記の翻訳を行うことが困難であることが分かる．一方，提案手法は用例間の類似度として品詞・活用形の情報のみを利用しているため，訓練データ中に出現しない表記の翻訳にも対応することができている．

7 おわりに

本論文では，対訳用例に基づく日本語機能表現表記の英訳手法を提案した．句に基づく統計的機械翻訳モデル Moses [3] との翻訳精度比較を行い，対訳用例を選定したテキストとは異なるジャンルのテキストにおいても，提案手法は比較的安定した翻訳性能を示すことを実証できた．今後は，機械学習の導入により，Moses のような統計的機械翻訳の手法と提案手法を併用することで，翻訳精度の向上を目指す．

参考文献

[1] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, H. Echizen-ya, T. Ehara, and S. Shimohata. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proc. 8th NTCIR Workshop Meeting*, pp. 371–376, 2010.

[2] グループ・ジャマシイ（編）. 教師と学習者のための日本語文型辞典. くろしお出版, 1998.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pp. 177–180, 2007.

[4] 文部科学省科学研究費特定領域研究「日本語コーパス」総括班: 特定領域研究「日本語コーパス」研究成果報告. 2011.

[5] 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.

[6] 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.