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*

島内 蘭†　阿部 佑亮†　鈴木 敬文†　宇津呂 武仁†　松吉 俊‡
筑波大学大学院 システム情報工学研究科†
奈良先端科学技術大学院大学 情報科学研究科‡

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．

日本語機能表現には，非常に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 [8]．本論文では，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 [10] を採用する．

文献 [8] 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 [7, 6] を利用して，日本語機能表現の日英翻訳を対象として，この SandGlass 翻訳方式を適用することにより，日本語機能表現の集約的な日英機械翻訳手法を提案している．本論文では，文献 [8] の成果をふまえて，日英対訳特許文を対象として，日本語機能表現の集約的英訳規則の作成および評価を行う．なお，翻訳規則作成のためには，目的言語側の訳が不可欠である．この際，目的言語側の訳が利用できない場合には翻訳作業を行う必要があり，文献 [8] では日常会話文に対して目的言語側の訳を作成したうえで，翻訳規則の作成を行っている．一方，本論文では，NTCIR-7 の特許翻訳タスク [1] で配布された 1,798,571 件の日英対訳特許文対を用いてフレーズテーブル [3] を学習し，日英対訳機能表現対を獲得するために用いた．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，翻訳規則作成が容易である点が大きな利点となる．日本語機能表現階層辞書 [7, 6] の 199 意味的等価クラスの中で，91 意味的等価クラスに属する日本語機能表現について，翻訳規則を作成し，その中の意味的等価クラス 12 個に属する日本語機能表現について評価を行なった結果，96.6%の正解率を得ることが出来た．

*Developing and Evaluating Rules for Translating Japanese Functional Expressions in Patent Documents into English through Canonical Expressions

†Ran Shimanouchi, Yusuke Abe, Takafumi Suzuki, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Suguru Matsuyoshi, Graduate School of Information Science, Nara Institute of Science and Technology

2　日本語機能表現

文献 [9] では文献 [4] で列挙された 125 個の見出し語だけでなく，その活用形を含めた 337 表現に対して，最大 50 文ずつの用例を文字列照合を用いて収集し，機能的な用法と自立的な用法の人手判定ラベルを付与した．また，文献 [7] 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」[1])．この辞書は文献 [9] 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ある．また，後に文献 [6] は，辞書内で言い換え可能な表現ごとに機能表現を分類し，言い換え可能な機能表現群ごとに意味的等価クラスラベルを付与した．

3　階層的日本語機能表現辞書

3.1　形態素に基づく階層構造

[7, 6] 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ている．これにより，図 1 に示すように，日本語機能表現の網羅的取り扱いが可能になった．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

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

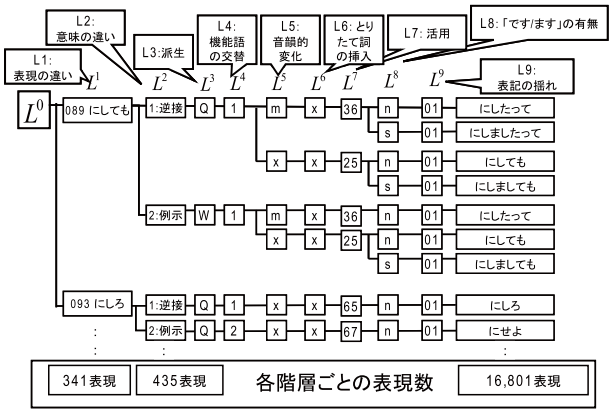

図 1: 形態に基づく階層構造

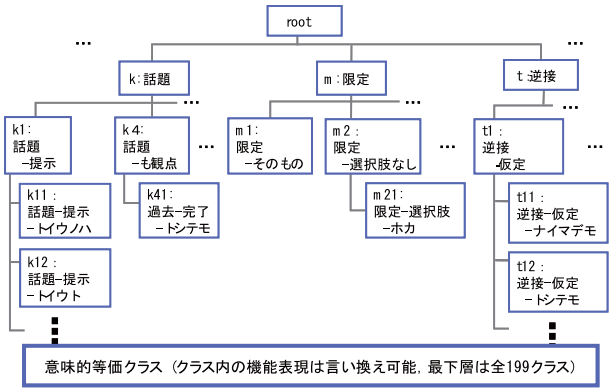

図 2: 意味的等価クラス

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．

3.2 意味的等価クラスに基づく階層構造

また，文献 [6] 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行った．図 2 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．

4 意味的等価クラスを用いた日本語機能表現の集約的英訳

文献 [8] では，「日本語機能表現一覧」の意味的等価クラス [6] の粒度を，日英翻訳用に再調整し，調整後

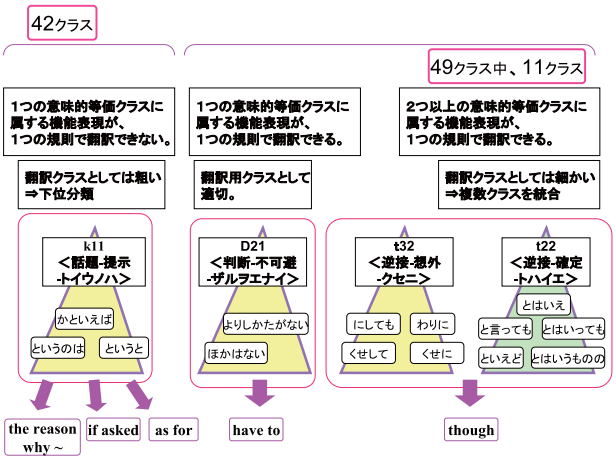

図 3: 文献 [8] における意味的等価クラスの粒度の再編

のクラスごとに翻訳規則を定めることにより，日本語機能表現を網羅的に集約的英訳する手法を提案している．文献 [8] においては，機能表現の用例文を集めるためのコーパスとしては，日本語文型辞典 [2] の電子テキスト版を用いた．この辞典は日本語学習者向けに機能表現の用例を約 8,000 文収録している．このコーパスにおいては，199 個の意味的等価クラスのうち，まず 91 クラスについて，1 クラス 5 文以上の例文を収集することができた．これらの 91 クラスについて，1 クラスから 5 文ずつ例文を抽出し，1 クラスあたり 1 つの翻訳規則で翻訳できるか否かの調査を行った．その結果，図 3 に示すように，下位分類が必要なクラスは 42 クラスであり，一方，1 クラスあたり 1 つの翻訳規則で翻訳可能なクラスは 49 クラスあり，49 クラス中の 11 クラスを計 5 規則に集約できることが分かった．

以上の文献 [8] の成果をふまえて，本論文では，1 クラスあたり 1 つの翻訳規則で翻訳可能な 49 個の意味的等価クラス，および，1 クラスあたり，複数の翻訳規則が必要な 42 個の意味的等価クラスの，計 91 個の意味的等価クラスを対象として，集約的英訳規則の作成を行う．

5 対訳特許テキストを利用した集約英訳規則の獲得

日本語機能表現の集約的英訳規則の獲得手順を図 4 に示す．

まず，日英対訳特許文対に対して，句に基づく統計的機械翻訳モデル [3] のツールキットである Moses を適用することにより，句の日英対応及びその確率を記

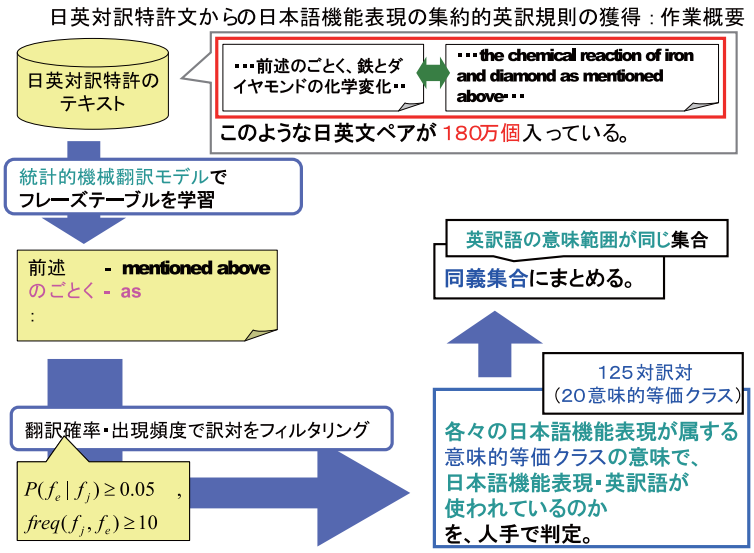

図 4: 日英対訳特許文からの日本語機能表現の集約的英訳規則獲得手順

載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から，大規模日本語機能表現階層辞書「つつじ」 [7] に収録されている機能表現のエントリを抽出する[2]．次に，これらの日本語機能表現および英訳語のエントリのうち，4 節で述べた 91 個の意味的等価クラスに含まれる日本語機能表現の表記，および，語義に該当する組のみを抽出した．その結果，20 個の意味的等価クラスに含まれる日本語機能表現を含む，125 組の「日本語機能表現-英訳語」組が抽出された．次に，125 組の「日本語機能表現-英訳語」組を，その日本語機能表現が属する 20 個の意味的等価クラスに分割した．最後に，20 個の意味的等価クラスの各々において，英訳語の意味・用法が同義となる「同義集合」へのまとめ上げをおこなった．

以上の手順の結果，20 個の意味的等価クラスのうち，16 クラスについては，「同義集合」の数は 1 つとなったが，残りの 4 クラスについては，1 クラス中の「同義集合」がそれぞれ 2 つに分割された．また，これらの 125 組の「日本語機能表現-英訳語」組に含まれる日本語機能表現の種類数は 57 表現であった．以上の集計結果を表 1 に示す．

このように，本論文の方式により，57 種類の日本語機能表現の英訳規則を 24 個に集約することができた．

表 1: 集約的英訳規則数および「日本語機能表現-英訳語」組数

	意味的等価クラス中の「同義集合」の数		
	1	2	合計
意味的等価クラス数	16	4	20
「同義集合」数	16	8	24
日本語機能表現の数	39	18	57
「日本語機能表現-英訳語」組数	80	45	125

6 集約的英訳規則の評価

集約的英訳規則の評価においては，5 節における集約的英訳規則獲得において，1 つの意味的等価クラスに対して，英訳語の「同義集合」が 1 つだけ存在する 16 個の意味的等価クラスを対象とした．そして，各意味的等価クラスにおいて，以下の全ての条件を満たす日本語機能表現を評価対象とした．

- 大規模的階層機能表現辞書「つつじ」 [7] において，意味的に多義性がない．
- 新聞記事 1 年分において 50 回以上の頻度で出現する．
- 機能表現表記について，新聞記事 1 年分から収集した用例に対して，機能的用法・内容的用法の間の用法判定 [9] を人手で行った結果，9 割以上が機能的用法で使われている．

[2]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 2: 集約的英訳規則の評価

クラス名	評価文数	正解率 (%)
D11(判断-当為-ナケレバナラナイ)	39	100
R11(比況-比況-ミタイ)	15	100
b11(対象-関連-ニツイテ)	85	100
e11(起点-極端例-ヲハジメトシテ)	15	100
e13(起点-極端例-カラ)	48	100
f12(範囲-範囲-ニワタッテ)	90	98.9
n21(添加-一様-ニヨラズ)	9	100
p12(継起-般-テカラ)	72	100
t24(逆接-確定-モノノ)	35	57.1
u12(対比-般-カワリニ)	20	100
v11(付帯-続行-ママデ)	30	100
y51(否定-当然の否定-トハカギラナイ)	6	100
計	464	96.6

これらの条件に該当する日本語機能表現は 23 表現であった．これらの 23 表現を少なくとも 1 つ含む意味的等価クラスは，評価対象の 16 個の意味的等価クラスのうち 12 クラスであった．これらの 23 表現について，対訳特許文からは,「日本語機能表現-英訳語」の組が 66 組収集された．そして,「日本語機能表現-英訳語」の組の各々に対して，評価文を最大で 10 文収集し，全体としては合計で 464 文を評価対象とした．

そして，これらの 464 文に対し，以下の 2 つの条件を満たす場合に，翻訳規則による翻訳結果が正しい訳であると判定する．

- 対象とする日本語機能表現の語義が，評価対象とした意味的等価クラスの意味に該当する．
- 評価用対訳文中の英訳部分が翻訳規則中の英訳と同義である．

評価の結果，表 2 に示すように，96.6%の正解率を達成することができた．この結果から分かるように，2 クラスを除いて，残りの 10 クラスについては 100%の正解率を達成することができた．このことから，これらの 10 クラスにおいて評価対象となった日本語機能表現については，評価文中においても，作成した集約的英訳規則と同一の語義の用法であることが分かる．一方，評価結果において翻訳誤りを含む 2 クラスについては，評価文中において，評価対象となった日本語機能表現表記が内容的用法で用いられていた．これらの日本語機能表現表記は，新聞記事中において内容的用法として用いられる割合は 1 割以下であるが，特許文中においては，内容的用法の割合が 1 割よりも高い可能性がある．

7 関連研究

代表的表現への言い換えを介した機械翻訳の研究としては，内容語と口語的な機能表現を扱った文献 [10] が知られている．また，本論文と同様に,「機能表現一覧」 [7] の機能表現を対象として，代表的表現への言い換えを介して機械翻訳を行う手法の研究としては，日本語文型辞典 [2] 中の例文を対象とした集約的英訳についての研究事例 [8]，および，集約的中国語訳についての研究事例 [5] がある．

8 おわりに

本論文では，日英対訳特許テキスト，および，既存の大規模日本語機能表現階層辞書中における 199 個の意味的等価クラスを用いることにより，日本語機能表現を集約的に英訳する翻訳規則を獲得する手法を適用した．16 個の意味的等価クラスについて，各意味的等価クラスにおける英訳語の「同義集合」が 1 つだけとなり，1 つのクラスあたり，1 つの英訳規則に集約可能であった．また，16 クラスのうち，12 クラスに属する日本語機能表現について，集約的英訳規則の評価を行なったところ，96.6% の正解率を達成することができた．今後は，他ジャンルの文書に対して本手法を適用し，集約的英訳規則の獲得および評価を行う．

参考文献

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

[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, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pp. 127–133, 2003.

[4] 国立国語研究所: 現代語複合辞用例集. 2001.

[5] 劉颯, 長坂泰治, 宇津呂武仁, 松吉俊. 意味的等価クラスを用いた日本語機能表現の集約的日中翻訳規則の作成と分析. 言語処理学会第 16 回年次大会論文集, pp. 194–197, 2010.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.