用例に基づいて英語前置詞の訳し分けを行う英日翻訳システム

5B−1

土井 伸一* 隅田 英一郎** 飯田 仁**
*NEC C&C 情報研究所 **ATR 自動翻訳電話研究所

1. はじめに

従来からのルールベースの翻訳手法に対して、近年、用例ベースの翻訳手法が提案されている。筆者らも既に、この手法が日本語助詞の英訳や英語前置詞の日本語訳に有効であることを示している [1, 2]。また最近では、より有効かつ現実的なシステムの構築を目指して、用例ベースと他の手法を組み合わせた手法も検討されている [3, 4, 5]。

筆者らは今回、このようなハイブリッドシステムの一つとして、基本的な解析・生成はルールベースで行い、英語前置詞の翻訳(係り先選択・訳語選択)は用例ベースで行う英日翻訳システム **REALIST**(Rule+Example=A LinguIST) を作成した。英語解析において、前置詞句が名詞句 / 動詞句に係るルールが適用されると、シソーラスを用いて入力と最も類似した用例を検索し、前置詞句の係り先と訳語を決定する。9種の前置詞に対して合計3000強の英日対訳用例を準備し、ATR会話コーパス約300文を対象とした実験で、概ね良好な翻訳結果が得られた。以下、システムの機能と構成、翻訳結果について報告する。なお、前置詞句の係り先選択の詳細は別稿 [6] で報告する。

2. 用例ベース翻訳とルールベース翻訳との融合

用例ベース翻訳とは、対訳用例をあらかじめ収集しておき、入力に対して、最も類似した用例を検索してその形式を模倣して翻訳する手法である。改良容易性、訳文の自然さ、ロバスト性等の特徴があり、ルールベース翻訳における「規則や辞書の作成・改良に多大なコストを要する」という問題点を克服するものとして注目されている。

しかしながら用例のみに基づく処理には、「様々な解析レベルでの用例を必要とする」「類似度計算のために入力表現から用例と同一の言語単位を切り出す必要がある」等、克服すべき課題も多い。またルールベース翻訳においても、そこで用いる知識はあらかじめ集積した用例から抽出したものであり、類似検索の処理を除けば、用例を参考にするという発想自体は自然である(例えば村木 [7] では、2概念間に成立する概念関係を保持する知識ベースを用いた機械翻訳が提唱されているが、ここでの概念関係は、収集した用例について各概念をシソーラスにより抽象化したものに相当する)。従って各々の手法は本質的に相容れないものではない。課題は、いかにして少数の規則によって処理できる言語現象と大量のデータに基づいて処理すべき言語現象を切り分けて、両手法を組み合わせるかにある。

筆者らの作成した英日翻訳システム **REALIST** は、基本処理はルールベースで、前置詞の翻訳は用例ベースで行うことを特徴とする。一般に自由格として扱われる前置詞の翻訳には、係り先や前置詞句中の単語の意味素性が複雑に関与するため、少数の規則での記述は困難であり、用例ベース翻訳が有効である。しかしながら、文中の単語すべてが前置詞句の係り先となれるわけではなく、係り先候補の抽出にはルールベース翻訳が有効である。このシステムは、自由格処理部として用例ベース翻訳関数をサブルーチンの形で組み込んだ翻訳システムと位置付けられる。

Example-Based Translation of English Prepositions
Shinichi DOI*, Eiichiro SUMITA**, Hitoshi IIDA**
*NEC C&C Information Technology Research Laboratories
**ATR Interpreting Telephony Research Laboratories

3. 用例に基づいた英語前置詞の翻訳

3.1. 用例ベース翻訳機能概要

英日翻訳システム **REALIST** の用例ベース翻訳機能の概要は以下の通りである(詳細は [2] 参照)。

翻訳対象 at, by, from, for, in, of, on, to, with の9種の前置詞。国際会議に関連する対話を対象としたATRコーパス異なり語約7万語の中で上位56位までに登場する高頻度のものを選択した。

用例(図1参照) 対象となる英語の前置詞句について、【係り先(名詞 / 動詞) 前置詞 前置詞句の名詞 head】の形の三つ組と、前置詞の訳語である日本語助詞(省略はZEROと記述)を対応させたもの。動詞については活用形情報(i:ing形, p:過去分詞, e:その他)も記述。9種の前置詞に対し合計3000強の用例を準備した。

シソーラス LEXICONの体系に準拠して作成した。

類似度計算 入力と用例との距離は、単語間の距離(シソーラス上で計算。現在、活用形情報は使用していない)と各単語の重みの積の総和で計算する。

例1 英語原文 The trip to Kyoto and the Ikebana class, right?
日本語文 京都旅行と生け花教室ですね。
抽出用例 (trip) to (kyoto) → (N2 ZERO N1)

例2 英語原文 Should I send it in cash?
日本語文 現金でお送りすれば、いいんですか?
抽出用例 (send e) in (cash) → (N2 で)

図1 英日対訳用例

3.2. 用例ベース翻訳機能を統合した翻訳アルゴリズム

英日翻訳システム **REALIST** の構成を図2に示す。この図に従い本システムの翻訳アルゴリズムを説明する。

1. 英語解析部は、単一化に基くボトムアップ横型のチャートパーザーにより、入力英文の解析を行う。前置詞句が名詞句 / 動詞句に係るルールの適用時に入力文から【係り先 前置詞 前置詞句の名詞 head】の形の三つ組を抽出し、用例ベース翻訳関数に送る(横型解析により可能性のある係り先をすべて抽出する)。
2. 用例ベース翻訳関数は、この三つ組と用例との距離を計算し、最類似(距離最小)用例との距離、最類似用例の頻度、前置詞に対する訳語を返す。
3. 英語解析部は、この距離・頻度を各解析結果の評価点の一つとして用いる。必須格の有無等から計算される評価点を併せて、最も高い評価点を持つ解析結果を選択し、対応する意味表現を出力する。
訳語は意味表現を介して日本語生成部に伝達する。

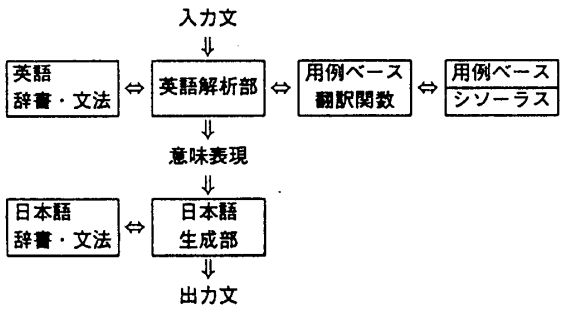

図2　英日翻訳システム REALIST 構成図

4. 日本語生成部では、意味表現から日本語文を生成する。意味表現上で前置詞に対する訳語が指定されている場合には、その訳語を出力する。

なお今回のシステムでは、用例ベース翻訳機能の評価を行う目的もあり、9種の対象前置詞については、at all, in front of 等の熟語を除き、すべてを用例ベース翻訳の対象とした。またこれらの前置詞句の係り先・訳語の決定には、非交差条件等の基本的な制約と用例ベースに基づく情報のみを用いた。これはすなわち、前置詞句に関する情報は動詞の辞書には記述せず、すべて自由格として扱ったことになる。さらに、係り先と前置詞句との物理的な距離(両者の間に存在する単語数)等の情報は使用していない。また生成の際も、「同一助詞の連続を制限する」等の制約を課すことはせず、用例ベース指定の訳語を生成する。

[正解例1] 入力英文 It is written in the notice.
日本語訳 通知に書かれています。
関数入力 (write p) in (notice)
検索用例 (write p) in (announcement) → (N2 に)

[正解例2] 入力英文 Also, please write your name in Romaji.
日本語訳 ローマ字でお名前もお書き下さい。
関数入力 (write e) in (romaji), (name) in (romaji)
検索用例 (write e) in (japanese katakana) → (N2 で)

[係り先不正解例] 入力英文 This will be my first visit to Japan.
日本語訳 これは日本に最初の訪問です。
関数入力 (be e) to (japan), (visit) to (japan)
検索用例 (be p) to (japan) → (N2 に)
I have not been to Japan for many years.
→ 私はここ何年も日本に行っておりません。

[訳語不正解例] 入力英文 Take the subway to Kitaoji Station.
日本語訳 地下鉄に北大路駅に乗って下さい。
関数入力 (take e) to (kitaoji station), …
検索用例 (get e) to (station) → (N2 に)

図3　英日翻訳システム REALIST による翻訳例

表1　翻訳結果の評価

係り先	訳語	
曖昧性なし	正解	31
曖昧性なし	不正解	1
正解	正解	90
正解	不正解	11
不正解	−	15

係り先の正解率
$$\frac{90+11}{90+11+15} = 87.1\%$$
訳語の正解率
$$\frac{31+90}{31+1+90+11} = 91.0\%$$

4. 翻訳結果とその評価

ATR会話コーパス約300文を対象とした翻訳実験を行い、翻訳結果に対して、用例ベース翻訳による前置詞句の係り先選択・訳語選択の評価を行った。翻訳例を図3に、全前置詞句(148例)に対する評価を表1に示す。

5. 今後の課題

上述のように、用例ベースに基づく前置詞句の翻訳について、係り先選択・訳語選択ともに90%程度の正解率が得られた。しかしながら用例ベース翻訳には様々な課題があることが指摘されている[4, 8]。今回のシステム作成でも、現在の形式の用例ベース翻訳機能の課題、もしくは用例ベースとルールベースの機能分担に関する課題がいくつか明らかになった。ここでは項目だけを列挙する。

- 動詞の活用形情報(記述済み)の利用。
- 三つ組以外の要素(修飾語、他の格要素等)の利用。
- 同一の三つ組での、前置詞句が必須格か自由格かによる訳し分け。
  ex.) I looked him at the door. = ～のところで
- 特殊な訳語を持つ必須格を用例として登録するときの、他への悪影響の可能性。
  ex.) major in = ～を専攻する
- 多義語が原因で発生する誤った推論。
  「多義語のシソーラスへの登録方法、用例として記述する際に意味分類を行うか否か」等の課題。

6. おわりに

基本的な解析・生成をルールベースで、前置詞の翻訳を用例ベースで行う英日翻訳システム REALIST を作成した。会話文約300文を対象とした実験で概ね良好な翻訳結果が得られ、用例ベース翻訳機能の効果を確認できた。

今後は、ここで明らかになった課題を考察し、用例ベース・ルールベースを始めとする様々な手法の有効な融合法を検討していく。例えば、必須格はルールベースで処理し、そこで対象とならなかった自由格だけを用例ベースで処理する方法も考えられる。また、あわせて用例データベースの充実を図り、用例ベース機能を組み込んだ実用的な機械翻訳システムの開発を目指す。

参考文献

[1] Sumita, Iida "Example-Based Transfer of Japanese Adnominal Particles into English", IEICE TRANS. INF. & SYST., VOL. E75-D, NO.4 JULY 1992
[2] Sumita, Iida "Example-Based NLP Techniques -A Case Study of Machine Translation-", AAAI-92 Workshop "Statistically-Based NLP Techniques", AAAI Technical Report W-92-01, 1992
[3] Furuse, Iida "Cooperation between Transfer and Analysis in Example-Based Framework", COLING-92, 1992
[4] 土井, 村木 "事例・統計情報を利用したハイブリッドな意味選択・訳語選択法の実現", 人工知能学会第6回全国大会, 15-5, 1992
[5] 浦本 "制約と事例による優先度を組み合わせた英文の多義性の解消", 情報処理学会自然言語処理研究会, 90-9, 1992
[6] 隅田, 土井, 飯田, 山端 "用例に基づいて英語前置詞句の係り先決定を行う英日翻訳システム", 情処第46回全国大会, 5B-2, 1993
[7] 村木 "知識ベースと, 言語に独立の中間表現とを用いた日英機械翻訳システム", 日経エレクトロニクス, 1984.12.17
[8] 渡辺, 浦本 "Example-Based Machine Translation の問題点に関する考察", 情処第45回全国大会, 2E-9, 1992