分野適応型翻訳機構のための翻訳不良表現の自動抽出

7Q－8

山田節夫　中岩浩巳　小倉健太郎　池原悟
ＮＴＴ　情報通信網研究所

１．はじめに

従来，ルールベース型機械翻訳システムでは，ある翻訳対象文における訳文品質を向上させる場合，その対象分野では頻出するがシステムでは正しく翻訳できない表現（翻訳不良表現）の訳文品質を改善するために，辞書やルールを人手によりチューンする膨大な作業が必要であった．我々はこの機械翻訳システムの分野適応化の作業を自動化するための枠組みとして，原言語と目的言語からなる対訳コーパスを用いた分野適応型翻訳機構を提案した[1].

本稿では，原言語の文を翻訳システムで翻訳した訳（機械訳）とコーパスの模範訳（理想訳）の構文構造を比較することによって翻訳不良表現を抽出し，その表現を問題別に自動的に分類する方法について述べる．

２．構文木の比較

まず，機械訳と理想訳の構文木を比較する方法について述べる．木の違いを見るためには，木と木の距離を用いた方法が提案されている[2,3]．これは基本的に，一方の木の部分構造を入れ替えて他方の木と同一構造にするときの変更部分が多いほど，距離を大きくするという考えで距離を定義している．しかしこの定義のみでは，構文木のようにノードのラベルによってその重要度が異なる場合を扱うには不十分である．そこで，我々は構文木の部分構造の独立性（ある部分構造が他の部分構造へ及ぼす影響度）に着目して構文木と構文木の距離を定義した．

定義　構文木と構文木の距離
1）一方の構文木の部分構造を入れ替えて他方の構文木と同一構造にするときの変更部分が多いほど，距離を大きくする．
2）入れ替える要素（部分構造）の独立性が強いほど，距離を小さくする．

この距離の定義に基づいて言語現象を対応付けると図１のようになる．実用的翻訳システムでは，これらの言語現象によって翻訳に用いる辞書がそれぞれ異なる．単語レベルで訳が異なるものは単語辞書で，用言レベルで訳が異なるものは用言構造変換辞書（用言と格の関係が記述されている辞書）で，文レベルで訳が異なるものはルールでそれぞれ翻訳している．以下，単語辞書によって解決される事例を**距離小事例**，用言構造変換辞書によって解決される事例を**距離中事例**，ルールによって解決される事例を**距離大事例**，また理想訳と機械訳が一致した事例を**距離φ事例**と呼ぶ．この距離を利用することによって解決法別に事例を分類することができる．

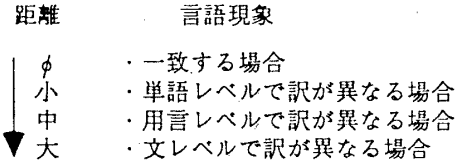

図１　言語現象と距離との対応

３．問題別事例抽出法

全ての文の基本である単文レベルの文について機械訳と理想訳の構文木を比較し，翻訳不良表現から距離φ事例，距離小事例，距離中事例，距離大事例を自動的に抽出する方法を考案した．この方法は，我々が開発している日英機械翻訳システムであるＡＬＴ－Ｊ／Ｅ[4]で採用した結合価文法に基づいた英語構文木を用いて検討した．事例は，述部同士及び格要素同士の対応の取れ具合によって表１のように分類される．この対応の取れ具合を以下の３つの場合に分けて考えた．

１．構文構造と訳が完全に一致
２．機械訳の構造のもとの原言語から理想訳の構造が生成可能
３．機械訳の構造のもとの原言語から理想訳の構造が生成不可能

１または２の場合，対応が取れているという．
述部の対応の取れ具合は，述部の対応が取れるか（「述部の対応」が"○"か），述部の訳が一致するか（「述部の訳」が"○"か）で評価する．表１において，「述部の対応」「述部の訳」が"○○"のときは１に，"○×"のときは２に，"××"のときは３に相当する．格要素の場合も同様に対応の取れ具合を決めているが，格要素は複数存在するので，全ての格要素の対応の取れ具合を考慮に入れなければならない．よって，全ての格要素の対応が取れているか（「格要素の全対応」が"○"か），全ての

表１　事例分類表

述部の対応	述部の訳	格要素の全対応	格要素の訳	事例
○	○	○	○	φ
○	○	○	×	小
○	○	×	×	大,(小)
○	×	○	○	中
○	×	○	×	中,小
○	×	×	×	大,(小)
×	×	○	○	中
×	×	○	×	中,小
×	×	×	×	大,(小)

Automatic Extraction of Badly Translated Expressions for a Domain Adapted Machine Translation System
Setsuo Yamada, Hiromi Nakaiwa, Kentaro Ogura and Satoru Ikehara.
NTT Network Information Systems Laboratories.

ＳＴＥＰ１：格要素に関する処理

↓ 理想訳の格要素の集合と機械訳の格要素の集合から対応の取れる格要素を抽出する

ＳＴＥＰ２：述部に関する処理

↓ 理想訳と機械訳の述部が一致するか辞書のエントリーと一致するかどうかを判断

ＳＴＥＰ３：距離別事例判定

ＳＴＥＰ１，２の結果を元に距離別事例を抽出し分類する

図２　距離別事例抽出アルゴリズム

訳が一致するか（「格要素の訳」が"○"か）で評価する．

事例は上記の対応の取れ具合によって分類される．大まかに言って，述部と格要素が共に１の場合は距離ϕ事例が，格要素が２の場合は距離小事例が抽出される．また，述部の訳が異なり全ての格要素の対応が取れた場合は距離中事例が抽出され，それ以外は距離大事例が抽出される．ただし，文全体としては距離大事例となってもその中に対応の取れる格要素があるとその部分だけ距離小事例となり得る．（これは表１中では（小）と表記している．）

格要素及び述部の対応の取れ具合によって，表１に基づき図２に示すアルゴリズムで距離ϕ事例，距離小事例，距離中事例，距離大事例を抽出する．

４．距離小事例抽出法

ここでは，距離小事例の抽出法（図２のＳＴＥＰ１）について具体的に考察する．

理想訳の格要素の集合と機械訳の格要素の集合から対応の取れる格要素を抽出し，そこから距離小事例を抽出する．１つの格要素が複数の格要素と対応が取れた場合，どの格要素同士が最も対応が取れているかを判断するため，表２に示すように構造と訳の２つの観点から，その対応の取れ具合を４種類に分けた．表２の中で，構造が「一致」とは機械の処理の都合によって構造が異なってしまった場合を含めて構造が一致することを示し，「他訳語候補」とは機械訳を訳出するときに機械翻訳システムが選ばなかった他の訳語候補を示す．

格要素の対応の取り方を図３の例で説明する．"装置"

日本語：本装置

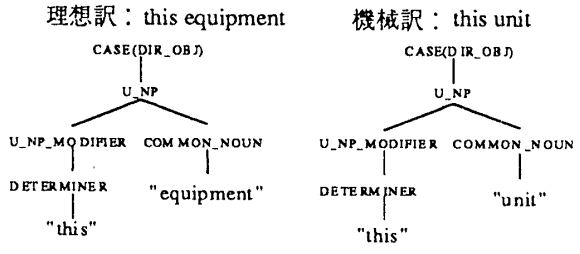

図３　格要素の対応例

表２　格要素対応表

訳＼構造	一致	不一致
一致	○	△
他訳語候補と一致	○	△
不一致	×	

の他訳語候補に"equipment"があった場合を考える．理想訳と機械訳の構造が一致しているので，表２の中の構造が「一致」に該当し，"装置"の他訳語候補に"equipment"があるので，「他訳語候補と一致」に該当する．よって両者の対応の取れ具合は"○"となる．

実際にＳＴＥＰ１についてのプロトタイプを作成し，分野固有の表現が多くその表現が比較的閉じている技術マニュアルから単文レベルの文（合計５７７文）を選び再現率（人手で対応の取れる格要素数に対するプロトタイプの割合）によりプロトタイプを評価した．その結果，45.2%の再現率が得られた．この再現率では十分とは言えないので，再現率を上げるための以下の案を考察中である．これらがすべて実現されると再現率が83.8%となることが見込まれている．

・システム辞書だけでなく一般の電子辞書も利用し，他訳語候補を探す．
・機械訳は見訳出１単語で理想訳は格要素１つの場合，それらの対応を取る．
・語尾の変形（派生語，ing形等）の違いは，英語辞書の情報を利用して対応を取る．

５．おわりに

本稿では，機械訳と理想訳を構造と訳の２つの観点で比較することによって，問題別に事例を４種類に自動的に分ける方法について述べた．その事例の中で，特に距離小事例について詳しく考察し，それらを抽出するプロトタイプを作成し評価した．その結果，45.2%の再現率が得られた．

今後は，再現率の向上策を検討すると共に，距離小事例から抽出された単語の登録方法を検討する．また，距離中事例，距離大事例についても検討する予定である．

参考文献

［１］中岩，山田，池原：対訳コーパスを用いた分野適応型翻訳機構，４８情処全大，7Q-07，1994
［２］Wilhelm, R.：A Modified Tree-to-Tree Correction Problem, Information Processing Letters, Vol. 12, pp. 127-132, 1981
［３］田中：構造をもつものの距離と類似度，情報処理，Vol. 31, No. 9, pp. 1270-1279, 1990
［４］池原他：言語における話者の認識と多段翻訳方式，情処論，28, No. 12, 1987