E-012

日中機械翻訳のためのスーパー関数抽出における対訳辞書自動構築

Automatic Construction of Japanese-Chinese Word Dictionary on Super-Function Extraction for Japanese-Chinese Machine Translation

李　楊 [1]　松本　和幸 [2]　北　研二 [2]　任　福継 [2]
Yang Li　Kazuyuki Matsumoto　Kenji Kita　Fuji Ren

1. はじめに

パターンを用いた機械翻訳手法[1]や用例ベースの機械翻訳手法[2,3]には，２つの問題点がある．まず，多様な文に対応できる翻訳パターンを作成するために，膨大な量の対訳コーパスおよび辞書を必要とすることである．品質の高い対訳コーパスや辞書を大量に準備するコストは非常に高く，いかにしてコーパスや辞書を拡充するかということも研究課題となっている．

もう一つは，訳語推定の問題である．パターンを用いる手法では，パターン中の変数部分（主に名詞）にシソーラスで定義されている意味属性を割り当てて，その制約条件をもとに，訳語を決定する仕組みである．しかし，シソーラスに含まれない語に対応できないため，柔軟性に欠ける．

用例ベースの機械翻訳の先行研究として，スーパー関数(SF)に基づいた日英機械翻訳[1]の研究があるが，翻訳時に SF の定数部分と一致しない入力文に対応できない．つまり，変数となる箇所が少ないほど１つの SF が対応できる文のパターンが少なくなるという問題がある．

そこで，本研究ではスーパー関数を用いた日英機械翻訳の利点と問題点を考察し，スーパー関数に基づく機械翻訳を用いて日中機械翻訳システムの試作を行う．また，新たに提案した SF の変数の拡張についても述べる．提案手法では変数となる品詞の種類が多くなるため，訳語を適切に選択するために，参照する単語辞書の自動作成手法を提案する．最後に，提案手法による自動翻訳実験を行い，スーパー関数に基づく日中機械翻訳の精度を向上するには，対象とする例文に応じた対訳単語辞書が必要であることを明らかにする．

2. スーパー関数に基づいた機械翻訳

2.1 従来の SF に基づいた日英機械翻訳

日英機械翻訳における SF は日本語と英語の対応を表す関数であり，名詞を変数とし，その他の品詞は定数と定義したものである．形式上の記述として表すと，

[O_STRing] < < O_Noun>+ [O_STRing] > * →
Super-Function(T_STRing,T_Noun)
[ ] :ちょうど１回の意味
+: 1回以上を意味する　*: 0回以上を意味する
0：原言語　T：目標言語
STRing: 名詞以外の自然言語文字列（定数）
O_STRing: 原言語文字列　T_STRing: 目標言語文字列
Noun: 名詞（変数）　→: 対応を表す

更に変数（名詞）はX,定数（名詞以外）はSとする．これらの添え字として$_o$及び$_t$があり，それぞれ原言語と目標言語を表す．また文字のつながりをΠで表すと以下の式が成り立つ．

$$S_{o0}\Pi X_{oi}S_{oi} = S_{t0}\Pi X_{tj}S_{tj}$$

S_{oi} ： 原言語定数　S_{tj} ： 目標言語定数
X_{oi} ： 原言語変数　X_{tj} ： 目標言語変数

例）　$f_1(X_3)$ ： X_{j1}はX_{j2}までX_{j3}に乗った．
= X_{e1} took X_{e3} to X_{e2} .
J ：[彼] は[駅] まで[タクシー] に乗った．
E ：[he] took [a taxi] to [the station] .

3. SF に基づく日中機械翻訳

3.1 定数と変数の定義

まず，日本語の助詞と助動詞のみを定数と定義する．その理由として，

- 日本語の助詞は直接翻訳文の文構造に影響を与える
  例）
  日：彼<u>が</u>笑われた。
  中：他被嘲笑了。
  日：彼<u>に</u>笑われた。
  中：被他嘲笑了。
- 中国語に日本語助詞を対応する訳文は不規則
  例）日本語助詞「と」の場合
  日：私は松本<u>と</u>言います。
  中：我叫松本。
  この例中の「と」は翻訳時に中国語の訳は存在しない．
  日：私は松本さん<u>と</u>行きます。
  中：我<u>和</u>松本去。
  この例文中の「と」は中国語の「和」と翻訳される．
- 日本語の動詞の接尾助動詞によって否定文，受け身文になるが，中国語での表現は様々である．

表 1: 否定文の日中対象

日本語	中国語
私は<u>行かない</u>。	我<u>不去</u>。
私は彼と<u>行かない</u>。	我<u>不</u>和他<u>去</u>。
私は彼と行くと<u>思わなかった</u>。	我<u>没想到</u>和他去

表１における例のように，日本語の「V+ない」の表現は中国語に翻訳する時に様々な形になる．

そこで，助詞と助動詞以外の品詞のすべてを変数と定義する．以下に，例を示す．

$\mathbf{F_1}$ ： N_{j1}はN_{j2}までN_{j3}にV_{j1}た．
= N_{c1} V_{c1} N_{c3} 去了 N_{c2} .

[1] 徳島大学大学院先端技術科学教育部
[2] 徳島大学大学院ソシオテクノサイエンス研究部

日：[彼] は[駅] まで[タクシー] に[乗っ]た.
中：[他] [乗][出租车] 去了 [车站] .

3.2 辞書の自動構築

まず，日中機械翻訳が用いた SF は原言語と目標言語の文構造のまま変数と定義された品詞部分を入れ替えることで翻訳が行われるため，品詞の意味を正確に捉えれば，より高精度の翻訳結果が得られる．しかし，動詞などを翻訳する際，目的語となる名詞によって対訳語が異なることもある．例で説明すると，

例：飲む（動詞）
ジュースを飲む → 喝饮料
薬を飲む → 吃药

上の例で，名詞「ジュース」を目的語とする場合，「喝」（日本語は“飲む”の意味をする），そして名詞「薬」を目的語とする場合に，「吃」（日本語は“食べる”の意味）となる．このように，訳語の曖昧性を持つ動詞は，一般的な対訳辞書中に沢山存在しているため，訳語決定のため，目的語となる名詞との関係を記録する必要がある．

本研究では提案した単語辞書は，単に対訳意味を記録するのではなく，SF 抽出時に，対応する SF と SF 中の位置情報も登録されている．これらの情報を用い，品詞（変数）間の共起出現頻度計算し，翻訳時に正しく品詞の意味を捉えるための情報とする．

例えば，「私は薬を飲んだ。」という入力文の場合，以下の SF が選択される．

F_1 ：N_{j1}はN_{j2}をV_{j1}た。= N_{c1} V_{c1} N_{c2} 了.

そして，変数を訳語に変換する時に，

1. N_{j1}（私）→ N_{c1}（我）
2. N_{j2}（薬）→ N_{c2}（药）
3. V_{j1}（飲む）→ V_{c1}（吃）/V_{j1}（飲む）→V_{c1}（喝）

となる．ここで，変換 3 の際に分岐が発生するため，訳語選択の処理が必要となる．この時に単語辞書における対応する SF の情報が必要となる．

まず，変数間の訳語の個数を θ $(x_i|x_j)$ 表すと

$V_{c1}= V_{j1}\theta$ $(V_{c1}|x_j)$ ；(θ=1)

この例では $X_j=(N_{C1};N_{C2})$ で，$X_j=N_{C1}$ の時に V_{c1}=(吃,喝)，つまり θ>1 であり，この場合，対訳参照としない．$X_j=N_{C2}$ の時に θ=1 になり，対訳参照とする．すなわち，V_{c1}=吃が得られ，訳文は，以下の通りとなる．

出力：我吃药了。

4. 評価実験

SF の抽出数について従来手法と比較すると共に，提案手法による翻訳実験の結果について述べる．

4.1 SF 抽出結果の比較

表 2：従来の抽出手法との比較

	SF 抽出数
従来手法	23,055
提案手法	12,243

表 2 で示した結果は 30,000 文対の日中対訳コーパスからの SF 抽出結果である．従来手法では名詞のみを変数と定義していたが，提案手法は，助詞と助動詞の以外の品詞を変数と定義している．しかし，用いた対訳辞書における日中単語の対応とコーパス中の単語の対訳関係とが異なっている場合があるため，提案手法で抽出出来た SF のうち，定義の通りに抽出できたものは 596 個であった．対訳例で表すと，

例：
日：エスカレータはどこですか。
中：自动扶梯在哪里?
理想的な SF の抽出結果：
F ： N_{j1}はN_{j2}ですか。
= N_{c1} 在 N_{c2} ？
実際のSF抽出結果：
F ： N_{j1}はN_{j2}ですか。
= 自动 N_{c1} 在 N_{c2} ？

日本語単語「エスカレータ」の中国語訳は対訳辞書中では「扶梯」である．また，中国語の「自动扶梯」を形態素解析した結果は「自动」，「扶梯」の二つの単語に分割されてしまう．この結果，対訳単語の照合を行い，SF を抽出する際に「自动」という部分が定数部分として残ってしまう．しかし，実際，このように抽出された SF が翻訳結果に影響することはほとんどない．

上述の問題を解決すれば，SF のデータベースのサイズをさらに縮小することが期待できる．

4.2 翻訳実験結果

対訳単語辞書に登録された単語を用いて 100 文の日本語文を作成し，翻訳実験を行った結果，86％の正解率が得られた．100 文の日本語文を作成する際，SF 抽出元のコーパスには含まれない文を作成した（オープン実験）．

この実験結果から，提案手法により構築した対訳単語辞書は SF にとって未知の単語への対応と訳語の選択処理も可能なことを明らかにした．

5. まとめ

対訳辞書の構築を自動化する目的は，機械翻訳に必要なデータベースを縮小するだけではなく，より少ないデータ量で多くの文の翻訳に対応することである．実験結果から，日中機械翻訳のためのスーパー関数抽出における対訳辞書の自動構築は今後の SF に基づく機械翻訳の精度を向上するために必要であることを示した．

今後の課題は，対訳辞書の単語量を増やすために，関連語などの概念を導入すること，そして，構築した辞書を参照し，意味フレームの概念を導入することで，少数の SF でより多くの文に対応できる翻訳システムを構築することである．

参考文献

[1] 楊 鵬,村上 仁一,徳久 雅人,池原 悟.
結合価パターンを用いた日中機械翻訳システムの構築，情報処理学会研究報告，Vol.2008，No.4，pp.121-126，2008.

[2] 篠山 学，黒岩 眞吾，任 福継.
Super-Function に基づく日英機械翻訳における日付・時間表現の抽出，電気学会論文誌 C，Vol.128，No.8，pp.1342-1350，2008.

[3] 桂 康，松本 和幸，任 福継.
名詞にかかる形容詞を対象とした Super-Function の拡張，第 72 回情報処理学会全国大会講演論文集(2)，pp.489-490，2010.