構造変換対を用いた複合名詞の英日-日英翻訳

上野 圭介　　宮崎 正弘

新潟大学大学院 自然科学研究科
{ueno, miyazaki}@nlp.ie.niigata-u.ac.jp

1　はじめに

複合名詞は基本的な単語を組合わせて際限なく生成される．機械翻訳においてこのような複合名詞を予めすべて対訳対として変換辞書（対訳辞書）に収録しておくことは困難である．従来，多くの機械翻訳システムでは，頻出の複合名詞を対訳対辞書に収録する方法を採用しているが，これでは辞書に収録する対訳対の網羅性の点で問題がある．

本稿ではこの問題を解決するために，英/日複合名詞の構造を予め解析し，抽象化された複合名詞の木構造の対訳対を用意することで英日/日英間の双方向の翻訳を行なう手法を提案し，その有効性を論じる．

2　従来の翻訳方法の問題点

機械翻訳における複合名詞翻訳では，原言語と目的言語間で単語同士を単純に置き換えられない非線形な部分があるため，こなれた良質の訳文が得られないことが多い．

従来の多くの機械翻訳システムでは複合名詞の翻訳において，簡単な対訳パターンを用いた翻訳処理を備えているものもあるが，頻出の複合名詞の対訳対を辞書に収録することを原則としている．しかし，このようなすべての複合名詞を予め辞書に収録しておけないため，複合名詞の構成単語が 1 単語変わっただけでもそのような複合名詞の対訳対を新たに辞書に追加しなければならず，辞書の収録数が際限なく増大し，辞書の構築・維持・管理が困難となる，といったような問題が起きる．

Translation Methods from English to Japanese, and Japanese to English for the Compound Nouns, using Constructional Transformed Pairs
UENO Keisuke, MIYAZAKI Masahiro
Niigata University

3　複合名詞翻訳システムの概要

2 章で述べた問題を解決するものとして，抽象化された日本語/英語の構造変換対を用いたマッチングによる変換と，構造化ルールを用いた構造変換によるハイブリッドな複合名詞翻訳を行ない，頻度情報を用いた訳語選択法により適切な訳語を得る英日-日英複合名詞翻訳システム（図 1）を試作中である．

半自動処理で生成された構造変換対と英/日構造変換ルールを用いて英語/日本語複合名詞を翻訳する．処理の流れは以下の通りである．

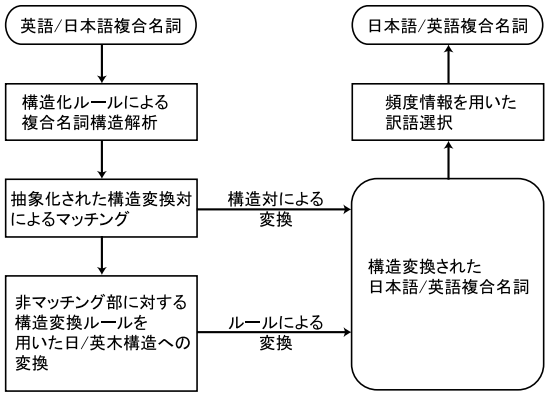

図 1：複合名詞翻訳システム“ *CompEdge* ”

本システムでは，構造変換対による翻訳を主な解析方法としており，補助的なものとして構造化ルールを用いた解析を位置付けている．木構造を用いた大域的翻訳である本手法には，以下の 3 つの利点がある．

1) 英日間で構成単語同士が対応しない日本語複合名詞を生成できる，単語間の要素合成法では複合名詞は変換が困難である．

2) 訳語選択で木構造に付随した訳語が参照できる．例えば図 2 の例では“ Conference ”という単語は“ 会議 ”と訳されるが，これを辞書引きすると“ 会議，協議会，相談，会談 ”のように複数の訳語があり，この中から最適な訳語選択をしなければならない．

3) 抽象化による網羅性の向上が期待できる．図 1 の変換対が登録されている場合，例えば“ 国際自然保護会議 ”のように複合名詞の構成要素の一部が異なる複合名詞でも正確な翻訳が可能となる．

4 言語データベース

本システムでは既存の日英対照辞書の他に，以下の 3 つのデータベースを準備している．

4.1 既出複合名詞 DB

翻訳実験を対象とした現存する組織名や役職などの日本語/英語の対で，入力された複合名詞にこれらが該当した場合，目的言語の訳を出す．

4.2 日英構造変換対（コーパス）

本システムの要となる大域的な構造変換のためのコーパスで，構造情報の他に，木構造に付随した訳語や英日間の単語のアライメントを伴っている．

図 2 は“ International Conference on Artificial Intelligence（国際人工知能会議）”を構造解析し，木構造を抽象化したもので，上部の木構造は模式的に表現されたものである．

ここで“ $1 ”は抽象化した部分単語列を表し，各単語の下方に付与されている数字はアライメントである．“ -1 ”は日本語との対応がない前置詞なので例外的に扱われている．

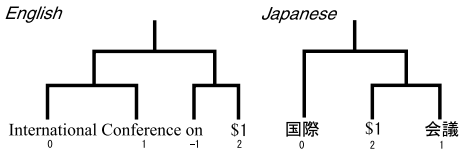

図 2： 構造変換対の例

4.3 訳語頻度データベース

訳語選択の際に，幾つかある候補の中から選出するための頻度情報を付加したものである．数値の算出法は後述する．

表 1: 訳語頻度データベースの例

種類	訳語, 頻度
Artificial	[人造, 3.13], [人工, 8.64], [模造, 1.42]
Intelligence	[知能, 1.23], [知性, 5.35], [知恵, 3.11], [情報, 0.82]

5 構造変換対と構造変換ルール

5.1 構造変換対の作成

構造変換対の作成は，基本的には半自動化で行なう．

まず，辞書や新聞等から収集して作成された既出複合名詞 DB と，複合名詞の構造を解析するための構造化ルール（5.2）を利用し構造解析処理を施して日英別個の木構造を作成する．次に，木構造で置換可能なワイルドカードとして有効な部分単語列を抽象化し，木構造マッチングにおいて変数として扱えるようにする．

以上のように生成されたデータと，人手で作成されたデータをマージさせ変換対を作成する．

本手法では構造変換に際して，入力された複合名詞の構造木とマッチングする方法を採用しているため，予めマッチングに使用するために構造解析をしておく必要がある．解析には，構造を含む生成規則を扱える拡張型のチャートパーザ [1] を用いている．

5.2 英語複合名詞の構造ルール

英語複合名詞の特徴として，主名詞（複合名詞の中で中心的な意味を持つ名詞）を修飾するための形容詞や前置詞句など名詞以外の品詞から構成されていることと，日本語複合名詞のように主名詞が常に最後尾に存在しないという点が挙げられる．

まず辞書データベースより複合名詞を構成する各単語の品詞を決定する．英語複合名詞は以下の 7 種類の品詞の単語で構成される．

表 2: 英語複合名詞における品詞の定義

品詞	表記	例
一般名詞	gn	conference, intelligence, etc.
固有名詞	pn	Japan, Miyazaki, etc.
形容詞	adj	artificial, primitive, etc.
接続詞	conj	and, &
決定詞	det	a, an, the
前置詞	pp	in, on, of, for, to
主形容詞	madj	international, central, etc.

ここで“ 主形容詞 ”とは，本来の形容詞と異なり複合名詞全体に係り得る形容詞を指し，現段階では上記の形容詞が該当し得る．これらの品詞情報を利用して英語複合名詞の木構造を作成する．

6 評価値を用いた訳語選択

辞書引きによって得られた訳語が複数存在する場合，その中から適切な訳語を選択する必要がある．そこで，辞書引きの結果，インターネットのサーチエンジンが返すヒット数，そして訳語頻度 DB の 頻度情報の，3 つのパラメータから算出した評価値（$GRADE$）によって訳語を選出する．

以下の図は計算の流れを示したものである．

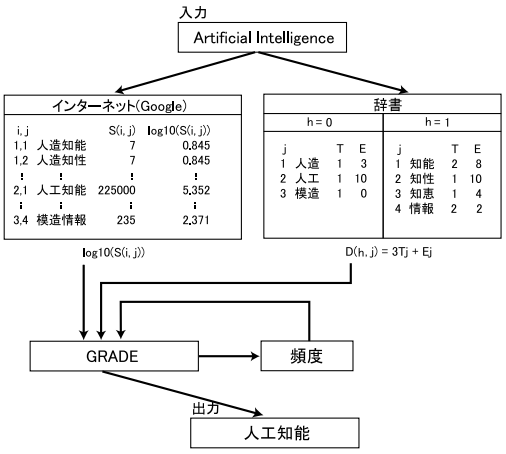

図 3：評価値の計算の流れ

6.1 英日-日英単語辞書を用いたダブル翻訳

まず，複数の単語辞書から各単語を翻訳する．その後，翻訳された単語をさらに辞書引きし，元の単語が訳語の中にあるかを調べる．

例えば“ intelligence ”では“ 知能，理解力，思考力 ”など，12 もの訳語がある．しかし，明らかに適切でない訳語もあり玉石混交である．この中から適切な単語を発見するために本システムではこれらの単語をさらに辞書で引き，その結果の中に元の単語，つまりここでは“ intelligence ”がある場合，その単語を有力な訳語としてみなす．この例では“ 知能，知性，知恵，情報 ”という 4 つの単語が候補して有力とされる．これらの情報を数値化し，パラメータとして加える．

6.2 サーチエンジンによる頻度付け

次に，上記の訳語から最も適切な組み合わせを選出するため，インターネットのサーチエンジンのヒット数を評価値のパラメータとして加え，最終的に評価値が最大の組み合わせを候補とする．サーチエンジンには世界で最も利用されている Google (http://www.google.com/) を利用する．

6.3 評価値の計算

上記 2 つの処理で得た評価値（解析 $GRADE$）と，訳語頻度 DB の頻度情報（頻度 $GRADE$）を組み合わせて最終的な $GRADE$ を計算する．

算出式は以下の通りである．

$$
\begin{aligned}
D_{(h,j)} &= 3T_j + E_j \\
A_{(i,j)} &= D_{(h,j)} + w \cdot \log_{10} S_{(i,j)} \\
{F_{(i,j)}}^{(t)} &= \sum^{k} \frac{A_{(i,j)}}{k} + w' \cdot {F_{(i,j)}}^{(t-1)} \\
{G_{(i,j)}}^{(t)} &= A_{(i,j)} + {F_{(i,j)}}^{(t)}
\end{aligned}
$$

ここで，辞書の訳語部と用例部に出現する訳語の回数をそれぞれ T と E，これらを計算した h 単語目の j 番目の辞書頻度を $D_{(h,j)}$，1 単語目の i 番目の訳語と 2 単語目の j 番目の訳語を組み合わせて検索した Google のヒット数を $S_{(i,j)}$，h 単語目の i または j の最大値を k，これらに重み w を付加して加えた解析 $GRADE$ を $A_{(i,j)}$，解析 $GRADE$ の平均と

$t-1$ 回目までの頻度 $GRADE$ を重み w' を付加した，解析 t 回目の頻度 $GRADE$ を ${F_{(i,j)}}^{(t)}$，そして解析 $GRADE$ と頻度 $GRADE$ の和である解析 t 回目の最終的な $GRADE$ を ${G_{(i,j)}}^{(t)}$ とする．

7　実験と評価

英語構造化ルールによる木構造作成と評価値による訳語選択の簡単な実験を行なった．

7.1　英日構造変換実験

実験には，翻訳者が使用する文献 [2] から組織名や役職等の 1093 の対訳対からランダムに抽出した 100 個の英語複合名詞を対象に日本語の複合名詞への構造変換実験を行なった．

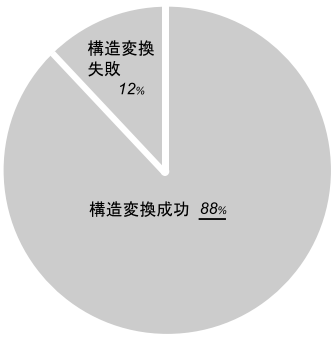

図 4：英日構造変換実験による定量評価の結果

これより，高精度で英日複合名詞翻訳が行なえる見通しを得た．

7.2　英日訳語選択の実験

資料 [3] から抜粋した小見出しの日英複合名詞 50 対を英語から日本語へ 6 章で述べた手法で訳語選択した．この対には例えば“ 人工知能, Artirifial Intelligence ”のように複合名詞として通常，辞書に収録されている専門用語等は除外している．本手法ではこのような複合名詞は辞書に登録されている語とマッチングすることで正確な訳語を選択することにしているので，本訳語選択のステップでは通常，専門用語等として辞書に収録されている複合名詞を翻訳する．結果は以下の通りである．

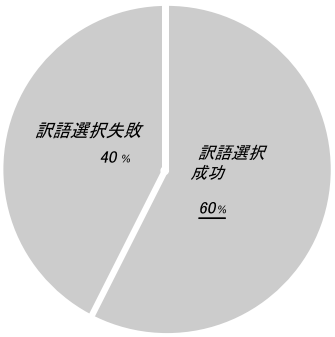

図 5：訳語選択による定量評価の結果

現時点ではまだ訳語選択の成功確率は 60 %であるが，重み付けの変化や細かな条件設定，もしくは辞書の追加などにより更に高精度なものにできつつある．

8　おわりに

本稿では大域的・構造的視点から，抽象化された構造変換対を用いたマッチング方式と構造変換ルールによる複合型の翻訳方式と，サーチエンジンと頻度情報による訳語選択の手法を提案し，その有効性を示した．本手法より，こなれた日英複合名詞表現の生成が可能になる．

今後の課題として，構造変換対の抽象化部の効率的なルールの拡充や，訳語選択で使用される評価値の更に高精度なパラメータを機械学習で調整することを考えている．

参考文献

[1] 川辺, 宮崎：構造を含む生成規則を扱える拡張型チャートパーザ, 言語処理学会第 11 回年次大会発表論文 (2005)

[2] 村田聖明：英文ライターのための和英翻訳ハンドブック, ジャパンタイムズ, (1990)

[3] 小島民雄他:情報･知識 imidas2000, 集英社, (2000)