自然な訳文を生成する機械翻訳システム実現のためのフレームワーク

A Framework for Generating Natural Sentences in Machine Translation Systems

武舎 広幸　　河村 政雄
Hiroyuki MUSHA and Masao KAWAMURA
マーリンアームズ株式会社
Marlin Arms Corporation
http://www.marlin-arms.co.jp/

機械翻訳システムによって生成されるほとんどの訳文は，原文の構造をほぼ直接的に反映する「直訳」である。読みやすい自然な訳文が生成されることは多くはなく，現在では，利用者もそれを期待してはいない。しかし，翻訳者が実務に用いているノウハウを利用すると，困難を伴う深い意味の解釈を行わなわなくても自然な訳文を生成できる場合も多い。本論文では，このようなノウハウをソフトウェアとして実現するための枠組みを提示するとともに，作成したプロトタイプを紹介する。

1　はじめに

筆者らは，長年にわたり機械翻訳ソフトウェアの開発を行うと共に，数多くの書籍や文書等の翻訳を行ってきた。また，近年翻訳教育に携わり，翻訳者を目指す人々に翻訳の際に必要な知識や技法などの教育を行っている。

人手による翻訳作業の際に，機械翻訳ソフトウェアの出力した訳文を用いる翻訳者の数は徐々に増えつつはあるものの，けっして多くはない。その理由は，機械翻訳システムによって生成されるほとんどの訳文は，たとえうまく構文が解析できたとしても，原文の構造をほぼ直接的に反映する「直訳」であり，そのままではいわゆる「下訳」として利用することもできず，自分で翻訳した方が速いと考える翻訳者が多いからである（翻訳者による機械翻訳システムの利用については[1]などを参照されたい）。

筆者らは人手による翻訳の経験，および翻訳教育の経験から，人間が行う翻訳作業にも，数多くの規則が存在し，翻訳者はその規則を機械的に適用する場合も多いことに気がついた。その後の調査により，こうした翻訳テクニックの規則は，とくに翻訳教育の現場で注目され，実際の翻訳教育に利用されていることも明らかになった（[2]など）。

そういった規則の中には，ソフトウェアとして実現するには，感覚的な判断基準に基づくため明確な手順として記述できず，当面実現不可能と思われるものも多いが，従来の機械翻訳システムの枠組みをより普遍的なものに拡張することにより，機械的に処理可能なものも多い。本論文ではこのような翻訳規則を具体例をとおして説明するとともに，そうした，いわば翻訳者のノウハウを明示的な規則として翻訳システムに組み込むためのフレームワークを提案する。

筆者らは，本論文で提案するフレームワークの一部を概念的に取り入れたシステムを構築し，小規模な辞書を用いたプロトタイプシステムを構築した。このプロトタイプシステムでは，頻繁に用いられる構文を含む文に対して，従来の翻訳システムでは不可能であった自然な訳文を生成することができる。

2　ソース言語辞書とターゲット言語辞書

まず，本論文で紹介するプロトタイプシステムがその基礎として用いている概念的な枠組みを説明する。図 1 は従来型の機械翻訳システムと本論文のシステムがベースとしているフレームワークの相違を示したものである。従来型の翻訳システムにおいては，辞書として，ソース言語（翻訳元の言語）の単語に対してターゲット言語（翻訳後の言語）ではどのような訳語が対応するかという，「対訳辞書」のみが用いられている。これに対して，本システムのもとになっているフレームワークでは，翻訳者が，対訳辞書の情報のみならずターゲット言語内部の語の関連も利用して，読者が違和感をもたない自然な訳文を生成するという事実に着目し，ターゲット言語内部の情報を記述する「ターゲット言語辞書」とソース言語内部の情報を記述する「ソース言語辞書」を用いる新しい枠組みを提案している。

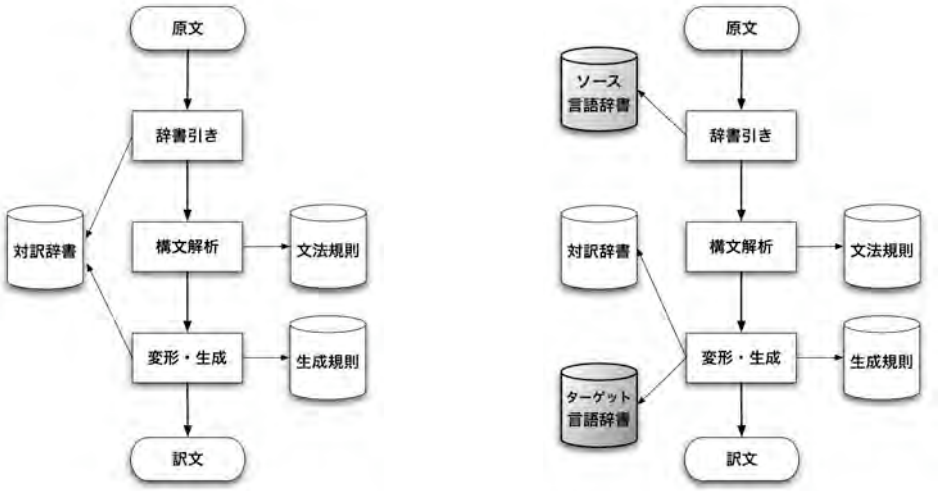

図１　従来の機械翻訳システムの構成（左）と本論文のフレームワーク（右）

この枠組みのうち，ターゲット言語辞書の有効性を，[2]から引用した次の例で説明する。

> Ignorance of foreign customs can result in unexpected misunderstandings. (1-1)

この例文は，翻訳者が訳す場合，たとえば次のような訳が考えられる。

> 外国の習慣を知らないと，思いがけない誤解を生ずることがある。 (1-2)

従来の機械翻訳システムにおいては，たとえば次のように訳される。この訳例はある翻訳システムのものであるが，他のシステムにおいてもほぼ同様の構文を持つ訳文となる。

> 外国の習慣についての不案内は，予想外の誤解に終わることがありえる。 (1-3)

従来システムの結果 (1-3) においては，原文で名詞であるignoranceがそのまま名詞で「不案内」と訳されているが，このような用法の場合，ignoranceのもつ動詞的な意味合いを日本語で的確に表現するためには，「知らないと」などと用言として訳すことが望まれる。このように，原文の品詞（たとえば名詞）を訳文では別の品詞（たとえば形容詞）に変える「品詞変換」の操作は，翻訳者による翻訳においては自然な訳文を出すためにきわめて頻繁に行われるが，従来の機械翻訳の枠組みでは不可能であった。

このような品詞変換を可能にするためには，図２に示すような，ターゲット言語内部の知識が必要になる。翻訳者の頭の中では，原文のignoranceに対応する訳語「無知」は，「不案内」「知らない」「知っている」などの単語と何らかの形でリンクされており，このリンクを何らかの判断基準を持って自由に行き来して，「無知」を「知らない」と変形しているわけである。

図２ ターゲット言語辞書の利用

この品詞変換に関して，より複雑な例を示す。次の例文 (2-1) は，たとえば (2-2) のように訳すことが望まれる。

> The document object gives you direct control over the cookies. (2-1)

> documentオブジェクトによりクッキーを直接制御できる。 (2-2)

しかし，従来の翻訳ソフトでは，たとえば次のように翻訳される。

> documentオブジェクトは，クッキーに対する直接の制御を与える。 (2-3)

(2-3)に出現する「…は直接の制御を与える」という表現は日本語として不自然である。感覚として不自然である理由を明示するのは容易ではないが，この例の場合，「与える」という動詞は「制御」という動作を表す名詞を目的語としてはとれないといった理由が考えられる。この例は，(2-2) のように「…により，直接制御できる」と訳すと自然であるが，このような訳を実現するためには，名詞controlを「制御できる」と動詞に変える必要がある。さらに，その名詞controlを修飾している形容詞のdirectは，「制御」が「制御する」に変わるのに伴って，副詞に変更される必要がある。すなわち，この例文では，名詞→動詞，形容詞→副詞という２つの品詞変換が行われている。また，原文では主語の役割をしている名詞objectは，「〜により」という送りが付加されて，少なくとも表層状は主語の役目を果たさなくなっている。

この例においては，２つの単語の品詞変換が行われているが，こうした変換は従来の枠組みではまったく不可能であった。

3. Word Worldモデル

現在実現しているシステムは英日の翻訳を行うシステムであり，上記のフレームワークに基づいているが，このシステムを拡張して多言語の翻訳システムへの応用を考えている。多言語を扱うシステムを考える場合，翻訳対象の言語対に依存した形で辞書を記述することは好ましくない。翻訳システムの辞書をソース言語辞書，ターゲット言語辞書，対訳辞書の３つに分けて考えると，対象言語の数が増えるごとに，扱わなければならない各辞書のペアが爆発的に増えてしまうのである。

しかし，これを図３のように，様々な言語の単語間にリンクが張られていると考えれば，各国語の単語の織りなす世界を統一的に扱えることがわかる。これを「Word World」モデルと呼ぶことにする。

図１に示したソース言語辞書，ターゲット言語辞書，対訳辞書のモデルは，人間の頭の中と比較すると，これにあまり似通ったモデルとは考えにくい。頭の中に入っている単語が言語ごとに別の「引き出し」に入っているというのは無理があると思われる。たとえば，日本語で「本」に相当するものの言語的な情報を人が頭の中で記憶しているモデルとしては，図３のような雰囲気のものが近いと考えられる。

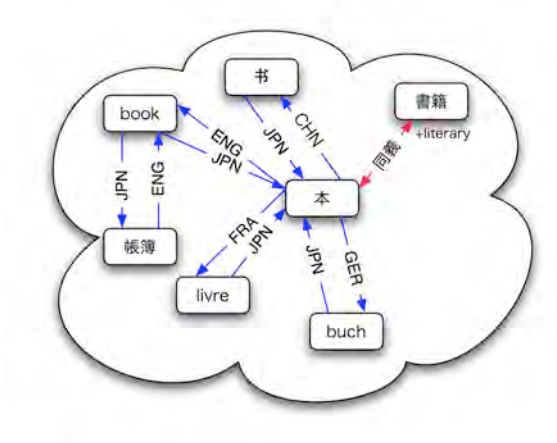

図３ Word Worldモデル

「Word World」においては，「本」という日本語の単語から，たとえば英語の「book」という単語へ，対応する単語であるという旨を示すリンクが張られている。同様に，（簡体字）中国語の「书」やフランス語の「livre」，ドイツ語の「buch」などへもリンクが張られている。英語の「book」からも当然「本」や「livre」などへのリンクがあるが，「book」には「帳簿」といった意味もあるので日本語の「帳簿」という単語へのリンクも張られている（同様にフランス語等の単語へのリンクもある）。日本語の「本」からは「書籍」という単語へもリンクがあり，これは「同義語」を表すものである。「本」と「書籍」を比較すると「書籍」の方が文語的雰囲気が強いので，図においてこれを「+literary」という「属性」を用いて表している。

このようなモデルを用いることにより，全世界に存在する単語を平等に扱ったシステムの構築が可能になる。人間の頭の中では，母国語のウェイトが高いが，母国語の単語も孤立しているわけではなく，ほかの言語の単語と並列に独立して扱われている。「Word World」のフレームワークは，この様子をより自然に表現していると考える。

4. システムの概要

4.1 本システムにおける翻訳過程のとらえ方

本論文に関して構築したシステムでは，上記のWord Worldモデルの前段階である，ソース言語辞書とターゲット言語辞書を用いるもフレームワークを念頭においたものである。実現までの時間を短くし，その有効性の検証を短期間で行うために，従来型のシステムをベースに構築し，考え方のみ上記のフレームワークを用いたものである。

機械翻訳システムは，その中身をブラックボックスとしてみれば，ソース言語の単語の並びというフラットな１次元的構造から，ターゲット言語の単語の並びという１次元的構造への変換を行うシステムである。このように入力と出力は一次元的な単語の列であるが，機械翻訳を行うためには，ソース言語の文の文法的構造を分析し，通常は構文木という２次元的な表現を用いてその文の構造を表し，ターゲット言語の構文構造を表現する構文木へと変形し，最後に再び１次元的な構造に戻して訳文を生成するのが普通である。

本論文のシステムにおいては，フラットな１次元的構造も，全体を統括する仮想的なノードを考えることにより，２次元的な木構造を持つものと考える。これにより，本システムでは翻訳の過程をソース言語の文字列からなる木構造をターゲット言語の文字列からなる木構造に変える過程と考える。つまり，本システムでは，すべての処理を，木構造の「変形」としてとらえるわけである。ただし，ここで言う「変形」には，一般的な意味における変形 — すなわち，ノードの接続関係の変化 — だけでなく，ノードへの何らかの情報（たとえば，品詞，訳語，意味的な属性など）の付加や，情報の削除も「変形」と表現することにする。

このように考えることにより，いわゆる辞書引き操作は，木構造のノードを構成する各単語に関する情報を辞書から得て，品詞などの情報を付加する「変形」操作としてとらえることができる。そして，辞書に記述された内容は，そういった変形操作を行うための「変形規則」を記述したものと考えることができる。

また，文法規則を参照しながら行う構文解析は，文法規則という「変形規則」を部分木に対して再帰的に適用していく過程と見ることができる。同様にして，従来の意味における構文変形規則は，木構造から別の木構造への変形の規則として記述される。また，最終的に訳文を生成する生成規則は，２次元的な木構造から，訳文文字列に不要な情報を落とす変形操作と見ることができる。

こうして，すべての処理過程を木構造の変形ととらえることにより，すべての処理を統一的に記述することができる。

4.2 翻訳例

次にあげる３つの文の翻訳過程について，従来のシステムとの相違を説明し，本システムで開発した実装を説明する。この３つの文の構文構造はいずれもきわめて似通っており，いくつかの単語が置き換わっているにすぎないものである。

> Her application of the theory to this case was in a sense quite natural. (3-1)

> My application of the theorem to this problem was quite natural. (3-2)

> His application of the theorem to this problem was not correct. (3-3)

これらは従来のシステムでは，たとえば次のように訳される。

> このケースへの理論の彼女の応用は，ある意味で全く自然だった。 (3-4)

> この問題への定理の私の適用は，全く自然だった。 (3-5)

> この問題への定理の彼の適用は，正しくなかった。 (3-6)

これに対して，本システムでは，たとえば次のように訳すことができる。

> 彼女が今回のケースにこの理論を当てはめるのはある意味ではきわめて自然なことであった。 (3-7)

> 私がこの問題にこの定理を当てはめるのはきわめて自然なことであった。(3-8)

> 彼がこの問題にこの定理を当てはめるのは正しくない。 (3-9)

このような訳文の生成を可能にするための辞書の記述を次に示す。図４は単語applicationに対する本システムの辞書の記述の例である。１行目の記述は単純な訳語の記述である。通常は，applicationに対しては，この訳語が選択される。本システムにおいては，木構造の該当するノードに「訳語」という情報を付加する「変形」規則として考える。これらの規則には条件

```
1 noun:{適用}
2 noun:[* [A +possessive][P "of" [O [A "the"] [P "to"]]]]
3   → [S [CL {のは} [V {当てはめる(S1)} [* <2 [1 {~が}]] [3 {~を} [4 {この}] [5 {に}]]]]]
```

図4 辞書の記述例

が記述されていないため，無条件で実行される。

これに対して，2-3行目の記述は，上記3-7から3-9の訳語を出すために用いられる定義（変形規則）である。→の左辺は「条件部」であり，右辺はその条件が満たされた場合に行われる変形を表現する「変形部」である

この変形規則の条件部は，名詞applicationの周囲のノードに関する条件を記述しており，次を要請している。

(A) ＊（application）ノードを所有代名詞（his, her, myなどpossesiveの属性を持つもの）が修飾している。

(B) さらに，＊ノードはofで始まる前置詞句で修飾されているが，その前置詞句の目的語（O）ノードは"the"と"to"で始まる前置詞句に修飾されている。

この変形規則の変形部では，次のような変形が行われる。

(a) ＊（application）ノードの上位に，節（CL）からなる主格（S）が新たに作られ，その主動詞として「当てはめる」を訳語としてもつV（動詞）ノードが新たに生成される（当てはめるに続く(S 1)は下 1 段活用を表す）。

(b) ＊ノードには左辺2番目の（P，of）ノードが合体され，左辺1番目（A，his）ノードには送り（助詞）として「が」が付加される。

(c) 左辺3番目（O，rule）ノードには送り「を」が付加される。

(d) 左辺4番目（A, the）ノードの訳語として「この」が指定される（theはデフォルトでは訳語を持たない）。

(e) 左辺5番目（P，to）の訳は「に」となる。これにより，「このケースに」の「に」が生成される。

この例でわかるように本システムの変形規則においては，次のような特徴がある。

- キーとなる部分だけをいくらでも詳しく，しかも（同じ文の範囲であれば）任意のノードに関して条件を記述できる
- 原文の品詞にかかわらず，新たなノードを作成し，そのノードにターゲット言語において自由な品詞を指定できる

5 まとめ

従来の翻訳システムで用いられている枠を超え，実務翻訳者が用いている翻訳技術を柔軟にシステムに取り込めるようにするためのフレームワークを提案した。また，このフレームワークに基づいてプロトタイプシステムを構築しその有効性を確認した。

謝辞

本論文の研究の一部は，独立行政法人情報処理推進機構（IPA）の2002年度および2003年度未踏ソフトウェア創造事業の補助の下に行われた。同事業の関係者，とくにプロジェクトマネージャとしてさまざまな面で支援していただいた日本エンジェルズ・インベストメント株式会社の紀信邦氏に感謝する。

参考文献

[1] 武舎広幸，機械翻訳しっかり入門，http://www.marlin-arms.co.jp/jpn/column/index.html。

[2] 安西徹雄，『翻訳英文法―訳し方のルール』，バベル・プレス，1986。