マルチ対訳コーパスを利用した多段用例翻訳方式

宮崎正弘
新潟大学　工学部　教授

１．はじめに

近年、コンピュータ技術と通信技術の進歩はめざましく、これらを統合して世界中のコンピュータを有機的に結合しようとするインターネットが急速な発展を遂げてきている。また、我が国の経済・社会などの諸活動も著しく国際化が進展しており、まさにボーダレスの様相を呈してきている。このような状況において、文化や言葉の壁が諸外国との種々の交流の大きな障壁となってきている。世界の言語の中で独自の位置を占める日本語の壁を取り除き、日本語と種々の言語（特に、インターネットや国際会議などで事実上の国際語となっている英語）間の高品質な機械翻訳を実現することが急務となってきている。

機械翻訳の研究は、コンピュータが出現してまもない１９５０年代から行われており、１９８０年代には日本でも日英や英日機械翻訳システムが製品化されているが、多くのユーザを満足するにたる十分な翻訳品質が得られていない。英仏、英独間など欧米言語間の機械翻訳に比べ、言語間で文法・語彙面だけでなく、言語表現の根底にある物事の捉え方や発想法が全く異なった日英間の機械翻訳は技術的にはるかに難しいものとなっている。また、言語には単語の語義から全体の表現の意味を合成できない慣用表現やある単語が特定の単語と表現の中に同時に出現しやすいという共起関係など汎用ルールでは扱いきれない様々な例外的・個別的表現が頻出する。このため、ルールベースの機械翻訳の限界が指摘されている。

このような限界を打破するものとして、用例ベースの機械翻訳（用例翻訳）が提案されている。[1] あまり英語に堪能でない日本人が日本語文を英訳する場合、翻訳しようとする文と似た翻訳例を見つけ、それを模倣して翻訳することが多い。用例翻訳は、このような翻訳のやり方をコンピュータ上で実現しようとするものであり、翻訳例を充実していくことにより、翻訳能力を向上させていくことが期待できる。用例翻訳については、既にいくつかの研究が行われているが、[2] 実用に十分耐えるシステムはいまだ実現されていない。対訳用例をいかに収集・蓄積して、対訳用例データベース（対訳コーパス）を構築するか、入力文と類似する用例を多くの用例の中からいかに効率よく探索するかなどが大きな課題となっている。特に、実用的な翻訳システムで対象とするような実際の文では、入力文を一つの用例でカバーできないことが多く、そのような場合、入力文をどのような部分に分割し、それぞれに対して、最も類似した用例をどのように探索するか、またそれら複数の用例をどのように組み合わせて入力文に対応する訳文を生成するかが大きな課題となっている。

本研究では、上記で述べたような用例翻訳の諸課題を解決するものとして、種々の観点から分類された日本語文と英語文の対訳コーパス（マルチ対訳コーパス）を利用し、一つの入力文に対して、マルチ対訳コーパス内の複数の用例を適用した日英間の用例翻訳（多段用例翻訳方式）を実現することを目指す。翻訳対象としては、表現が多様で、原言語と目的言語間での表現の差が大きく、１対１の対応づけが容易でない、会話文や電子メール文等のような人間の感情・意志などの情緒的表現に富む文をとりあげ、用例翻訳による高品質な機械翻訳の可能性を探る。

２．多段用例翻訳方式

用例翻訳において、一つの述語（用言）が含まれる単文だけでなく、重文や複文といった複数の述語（用言）が含まれる入力文を翻訳しようとする場合、通常、一つの入力文全体をカバーするような類似用例を対訳コーパスから検索できない。また、多様な英語表現に対応する名詞句や複合名詞などのような句を翻訳するには、名詞句や複合名詞の対訳コーパスの利用が有効である。

図１に示す”多段用例翻訳方式”では、用例翻訳に必要な対訳コーパスの量をあまり増やさずに重文や複文を翻訳対象とし、名詞句や複合名詞に対する質のよい翻訳を実現することを目指し、一つの入力文に対して、様々なレベルの対訳コーパスであるマルチ対訳コーパスを参照することによって多段階に翻訳を行う。辞（文）、詞（述語＜用言＞）、句、複合語のような文を構成する様々な階層に応じた日英の対訳用例を構文解析したうえで、構文解析済みの対訳コーパスとし

てデータベース化したマルチ対訳コーパスを構築しておく。

マルチ対訳コーパスを効果的に利用するため、それぞれのレベルに応じた類似評価を行う各種のモジュールが存在する。日本語の入力文は、まず形態素解析、構文解析が行われ、木構造の形式で翻訳部に受け渡される。翻訳部では、この木構造から上記のモジュールを呼び出し、これらのマルチ対訳コーパスから入力された日本語文に類似する様々な日本語類似表現素片を検索し、そこから類似表現素片とその差分である差分表現素片、その各々に対応する英語類似表現素片と差分表現素片を抽出する。

次に、それらを組み合わせて入力文に対応する英語文を生成する。ここで、ポインタで結ばれている素片と素片を結合する際、ただ単純につなげばよいのではなく、結合する素片によっては素片に対して、付加・変形・削除などの処理を行う必要がある。この処理は単文結合型英文生成で行われる。最後に、語尾変化処理などの形態素調整を行い、文法的に正しい英語文を生成する。

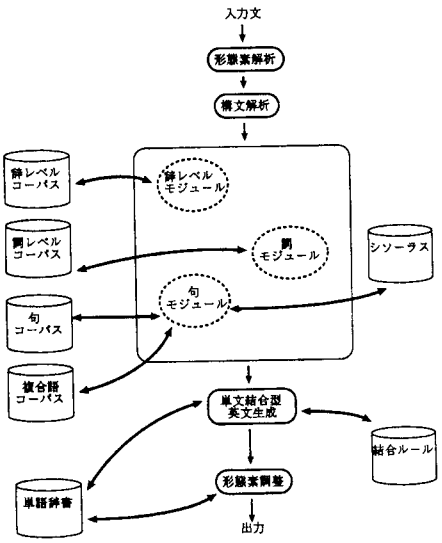

図１．多段用例翻訳システムの概要

３．辞（文）の翻訳

３．１　主体的表現と客体的表現の分離

人間の言語活動における過程的構造に着目した時枝誠記による言語過程説を発展的に継承した三浦つとむの日本語文法（三浦文法）によれば、[3),4)]一つの文は主体的表現と客体的表現から構成される。主体的表現は人間の感情、意志など主観的な判断を直接表現したものであり、客体的表現は対象を概念として捉えて表現したものである。日本語では、主体的表現の語（辞：助詞、助動詞、感動詞、接続詞など）と客体的表現の語（詞：体言、用言、副詞、連体詞など）がそれぞれ別の単語になっているため、主体的表現と客体的表現を簡単に分離できる。宮下眞二はこのような三浦の考えを英語文に適用し、独自の英語文法（宮下文法）を提案している。[5)]

このような考えに基づき、三浦文法をベースとした日本語形態素解析システム[6),7)]を構築すると共に、一般化ＬＲ法（富田法）を Prolog 上に実現した拡張型のＳＧＬＲパーザ[8)]である SGLR-plus[9)]上に、三浦文法による日本語文法、宮下文法による英語文法を補強項付きＤＣＧ形式で記述することによって、日本語文パーザ[10)]、英語文パーザ[11)]を構築した。これらのパーザによって、日本語文や英語文を構文解析し、主体的表現と客体的表現を分離することができる。本研究では、これらのパーザを入力文の構文解析や構文解析済みの対訳コーパス（日本語文・英語文）の構築に利用している。

３．２　辞（文）レベルの対訳コーパスの検索

人間の感情・意志などの情緒的表現に富む文では、主体的表現を担う辞が重要な役割を果たし、日英翻訳における目的言語である英語文の表現形式を決定づける。会話文や電子メール文等のような情緒的表現に富む文を対象とする用例翻訳を行うには、様々な主体的表現を含む文の対訳コーパス、すなわち辞（文）レベルの対訳コーパスを構築する必要がある。

日本語文パーザによって作成された、詞と辞を分離した木構造を、辞をキーとして上記の辞（文）レベルの対訳コーパスと照合し、入力された日本語文に類似する日本語類似表現素片を検索し、そこから類似表現素片とその差分である差分表現素片、その各々に対応する英語類似表現素片と差分表現素片を抽出する。

差分がある場合、差分表現素片もまた対訳コーパスと照合し、入力された差分表現素片に類似する日本語類似表現素片を検索し、そこから類似表現素片とその差分である差分表現素片、その各々に対応する英語類似表現素片と差分表現素片を抽出する。以上の処理を差分がなくなるまで繰り返す。ここで、照合すべき対訳コーパスの種別は、上位の日本語類似表現素片から差分表現素片への継承情報（６章参照）によって、辞（文）、詞（述語＜用言＞）、句、複合語の対訳コーパスのどれかに決定される。

例えば、入力文「新潟駅へ行く道を教えて下さいませんか」は、詞「新潟駅へ行く道を教える」と辞「依頼・疑問（丁重）」に分離される。次に、辞「依頼・疑問（丁重）」キーとして辞（文）の対訳コーパスと照合し、日本語類似表現素片「～へ行く道を教える＋依頼・疑問（丁重）」とその差分表現素片「新潟駅」、英語類似表現素片”Could you tell me the way to ～?”を得る。
次に、差分表現素片「新潟駅」を複合名詞の対訳コーパスと照合し、最終的に複合名詞「新潟駅」に対応する英語訳として”Niigata Station”を得る。

４．詞（用言）の翻訳

用言にはそれが要求する名詞と格助詞からなる格要素があり、用言とそれが支配する格要素は一つの意味のまとまりと考えられる格パターンを構成する。用例翻訳によって用言を翻訳するには、用言ごとに格パターンに相当する用例を収集・蓄積する必要がある。しかし、用例翻訳において、対訳コーパス量と翻訳品質の定量的関連についてはあまり明確になっていない。ここでは、格要素として多義名詞を含む詞（用言）のコーパスから類似文を正しく抽出し、動詞の訳語選択を行う方法について述べる。さらに、そのような処理を行う場合の対訳コーパス量と翻訳品質の関連を評価分析した結果を示し、対訳コーパス構築の指針を明らかにする。

４．１　詞（用言）の対訳コーパスの構築

用例翻訳において対訳コーパスから入力文と最も類似した文を抽出する際、名詞の類似度を判定するため名詞シソーラスを用いる。ここでは、名詞の意味の近さを判定するために、名詞意味属性体系[12]を用いる。そのため対訳コーパスの日本語側の格要素に意味属性を付加する。しかし、人手でコーパスの名詞の多義を絞り込んでおくことは、大量のコーパスを用意するという観点から問題となる。

そこで対訳コーパスを自動収集し、人手加工を行わず利用するため、複数の意味属性を持つ場合も、多義を絞り込まずに全ての意味属性を付加する。ここで、「金＞ gold、money、--」のように日本語側では多義がある名詞を英語の対訳から意味属性を絞り込める場合、共通の意味属性に絞り込んでおく。

４．２　類似文抽出法

入力文における格要素の字面・意味属性と、コーパスの格要素の字面・意味属性との類似度を調べて、入力文と最も類似したコーパスを抽出し、訳語選択を行う。類似文の検索は動詞をキーにして行う。意味属性間での類似度の評価法については、通常、シソーラス上で近いもの同士ほど、高い評価値を与える。しかし、シソーラスは分類の細かさが場所によって異なるので、ここでは単純にシソーラス上の距離を類似度とするのではなく深さを類似度とする。この評価を意味属性の多義を絞らずに全ての多義の組合せについて行い、最も良い評価点をその２名詞間の類似度とする。

以上の評価を、入力文のそれぞれの格要素とコーパスのそれぞれの格要素を比較して行う。さらに、格ごとに重み付けを行い、入力文とそれぞれのコーパスとの類似度の評価点を求める。その中で最も評価点が高いコーパスを抽出し、動詞の訳語選択を行う。

４．３　評価

多義動詞「とる」、「あがる」について訳語選択の評価を行った。格の重み付けは他動詞「とる」では「を格」に他の格の１０倍の重みを付けたが、自動詞「あがる」では全ての格を同じ重みとした。用例では対訳コーパスの名詞の多義を人手で絞り込んだ場合と、絞り込まなかった場合を比較した。対訳コーパスを用いる方法と格パターン（日本語語彙大系[12]のパターン対辞書）

を用いる方法との比較を表１に示す。対訳コーパス量に対する正解率を図２に示す。

対訳コーパスを自動収集することを前提に、名詞の多義を絞らずに類似度評価する方法を提案し、その評価を行った。本評価より以下のことが明らかとなった。

・格パターンを用いる方法より良い正解率が得られた。これは格パターンの網羅性がまだ十分でないためと考える。
・名詞の多義を絞り込んだ場合とそれほど変わらない正解率が得られた。
・動詞ごとの対訳コーパスの数が１００文を越えると、正解率は目立って上がらなかった。慣用的な表現を狙って記述しても正解率は目立っては上がらないためと考える。

表１．「とる」「あがる」の訳語選択結果

	とる	あがる
用例（多義の絞り込みなし）	84.0%	90.0%
用例（多義の絞り込みあり）	82.0%	86.7%
格パターン	70.0%	73.3%

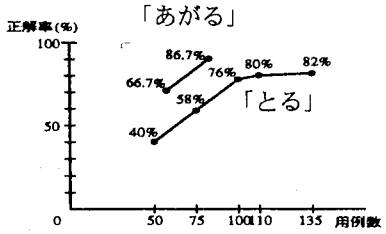

図２．対訳コーパス量に対する正解率
（「とる」「あがる」の訳語選択）

５．詞（体言）の翻訳

格要素を構成する詞（体言）には、１単語名詞である単純名詞の他に、複数の単純名詞や名詞相当の接辞が結合して作られる複合名詞、複数の名詞（単純名詞、複合名詞）が助詞「の、や、--」を介して結合して作られる名詞句がある。日本語文には、名詞と名詞を助詞「の」で結合した「ＮのＮ」という形式の名詞句や複合名詞が頻繁に出現する。ここでは、「ＮのＮ」型名詞句と複合名詞をとりあげ、用例に基づく翻訳法を提案し、評価実験によってその有効性を示すとともに、対訳コーパス構築の指針を明らかにする。

５．１　対訳コーパスを用いた「ＮのＮ」型名詞句の日英翻訳

「ＮのＮ」型名詞句は、その意味構造が多様であり、また英語との対応が１対１でないため翻訳規則を定式化することが困難である。そこで、用例翻訳が有効と考えられる。ここでは、「ＮのＮ」型の名詞句に対応する英語表現を７種類に分類し、それぞれの対訳コーパスを構築した。入力された「ＮのＮ」型名詞句と最も類似する名詞句を対訳コーパスから探索するとともに、対訳コーパスとの差分を抽出し、それに基づいて翻訳を行う。名詞の類似度の判定には、詞（用言）と同様の方法を用いるが、名詞自身が翻訳対象となっているため、名詞に多義がある場合、名詞の多義は対訳コーパス構築時に人手により絞り込んでおくこととした。

対訳コーパス量に対する正解率を図３に示す。本評価より以下のことが明らかとなった。

・対訳コーパス量を約５００とした場合、商用の翻訳ソフトより高い８８％の正解率が得られた。
・コーパス量を増やせば正解率は向上するが、処理速度が遅くなってしまう。今後、処理速度の向上について、検討する必要がある。

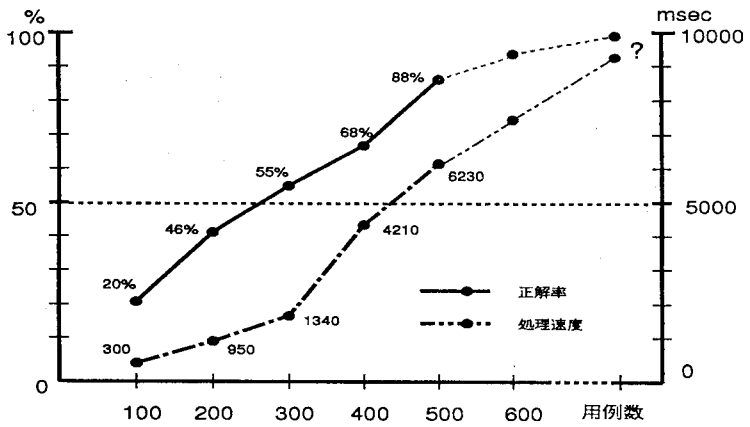

図３．「ＮのＮ」型名詞句の対訳コーパス量に対する正解率

５．２　対訳コーパスを用いた複合名詞の日英翻訳

日本語複合名詞は、造語力の強い漢字が結合して構成されることが多く、その組合せによって数限りなく生成されるため、そのすべての対訳パターンをあらかじめ用意しておくことは困難である。また、その対訳である英語表現には様々な表現があり、翻訳規則として定式化することが困難である。そこで、用例翻訳が有効と考えられる。ここでは、２単語より構成される複合名詞に対応する英語表現を５種類に分類し、それぞれの対訳コーパスを構築し、これを用いた用例翻訳を行う。

入力複合名詞が２単語の名詞相当語で構成されている場合、入力複合名詞と最も類似する複合名詞を対訳コーパスから探索するとともに、対訳コーパスとの差分を抽出し、それに基づいて翻訳する（局所的用例翻訳）。名詞の類似度の判定は「ＮのＮ」型名詞句と同様の方法を用いる。

入力複合名詞が３単語以上の名詞相当語で構成されている場合、入力複合名詞の構造解析[13]によって得られた木構造の中のより小さな部分構造（２単語の名詞相当語で構成される）を局所的用例翻訳によって翻訳し、その翻訳結果を部分構造の主名詞と部分構造を包含したより大きな構造の名詞とを局所的用例翻訳によって翻訳した結果に埋め込む。以上のように局所的用例翻訳を繰り返してより大きな構造の複合名詞全体を翻訳する。図４に本方法による翻訳例を示す。

構成単語数が２から８の複合名詞１００個を対象に翻訳実験を行った結果、以下のことが明らかとなった。

・対訳コーパス量を約３００とした場合、商用の翻訳ソフトより高い７７％の正解率が得られた。全体の９％を占める訳語選択による誤りをなくせば、正解率を８６％まで向上させられる。
・局所的用例翻訳を繰り返して適用することによって翻訳できない複合名詞には、３単語以上で構成された大域構造化された対訳コーパスによる用例翻訳が必要とされる。

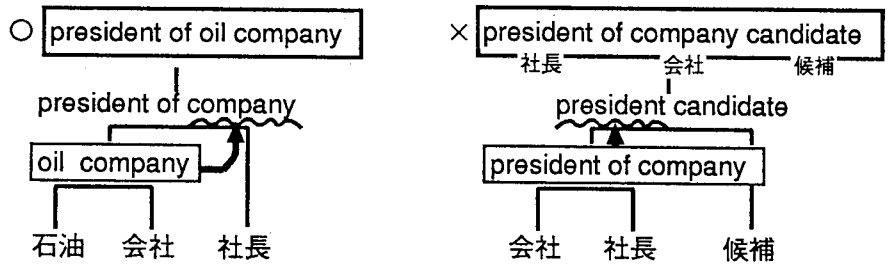

図４．対訳コーパスを用いた複合名詞（構成単語数３以上）の日英翻訳例

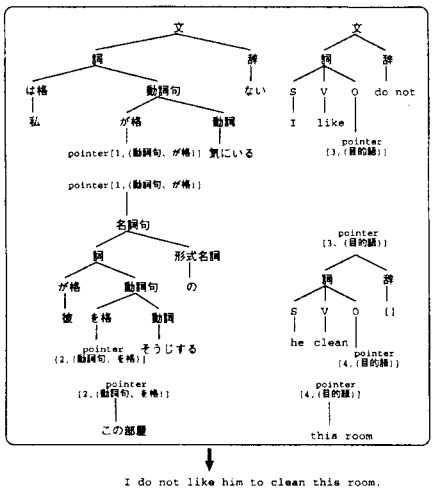

図５．単文結合型英文生成による処理例

６．単文結合型英文生成

前章までの処理によって最終的に複数の、日本語と英語のそれぞれの類似表現素片と差分表現

素片を得る。この複数の英語類似表現素片の木構造のポインタとポインタを結合して、一つの木構造を生成する。

ポインタとポインタをつなぐ時、単純に結合すればいいというわけではなく、木構造を変化させる必要がある。この木構造の処理をする際に継承情報を利用する。この処理にはルールを用いる。継承情報はコーパスにあらかじめ付与されている場合と、素片の組み合わせから抽出する場合がある。

例えば、図５の日本語木構造では、ポインタ１の下位の構造に形式名詞「の」が存在し、上位の構造の動詞句に「が格」でつながっているという継承情報をのせ、ポインタ２にはただ単に上位の構造の動詞句に「を格」でつながっているという情報にのせる。また、図５の英語木構造では、ポインタ３には動詞 like の目的語であるという情報をのせ、ポインタ４には動詞 clean の目的語であるという情報をのせる。ここから、「英文を生成するさいに、動詞 like は目的語に to 不定詞をとり、ポインタ３以下を to 不定詞化する」といった処理を行い、木構造を合成して英文を生成する。このように素片を結合するさいに特定の単語に固有の構造を作ることが多い。

７．おわりに

一つの文は主体的表現と客体的表現から構成され、人間の感情・意志などの情緒的表現に富む文では、主体的表現を担う辞が重要な役割を果たし、日英翻訳における目的言語である英語文の表現形式を決定づける。このような点に着目して、辞（文）、詞（述語＜用言＞）、句、複合語のような文を構成する様々な階層に応じた、構文解析済みの対訳コーパス（マルチ対訳コーパス）を利用し、一つの入力文に対してマルチ対訳コーパス内の複数の用例を適用した用例翻訳として多段用例翻訳方式を提案し、マルチ対訳コーパスの構築法について論じた。

現在、会話文や電子メール文等のような情緒的表現に富む文を対象にした、多段用例翻訳方式に基づいた日英間の用例翻訳システムのプロトタイプの試作を進めている。今後、マルチ対訳コーパスの充実を進めるとともに、翻訳品質や処理速度等の総合的評価を行い、本方式の有効性を検証していく予定である。

[謝辞]　新潟大学工学部宮崎研究室において、本研究に参加された大学院生・学部生（池田修一、富樫亮介、加藤朋幸、杉浦徹哉、室谷祐子、高沢恵美子、本間寛幸、武本裕の諸君）、および日本語語彙大系データ、対訳コーパス用原データなど本研究に必要な各種の言語データを提供して下さったNTTコミュニケーション科学基礎研究所の翻訳研究グループの関係各位に深謝する。

＜参 考 文 献＞

1) M.Nagao：A Framework of a Mechanical Translation between Japanese and English by Analogy principle，Artificial and Human Intelligence，North-Holland，PP.173-180（1984）
2) 佐藤理史：MBT2:実例に基づく翻訳における複数翻訳例の組合せ利用，人工知能学会誌，Vol.6，No.6，pp.861-871（1991）
3) 三浦つとむ：日本語とはどういう言語か，講談社学術文庫（1976）
4) 宮崎，池原，白井：言語の過程的構造と自然言語処理，「自然言語処理の新しい応用」シンポジウム論文集，pp.60-69（1992）
5) 宮下眞二：英語とはどういう言語か，季節社（1985）
6) 宮崎，白井，池原：言語過程説に基づく日本語品詞の体系化とその効用，自然言語処理，Vol.2，No.3，pp.3-25（1995）
7) 高橋，佐野，宍倉，前川，宮崎：頑健性を目指した日本語形態素解析システムの試作，「自然言語処理における実働」シンポジウム論文集，pp.1-8（1993）
8) 沼崎，田中：SGLR: 逐次型一般化 LR パーザの Prolog による実現，情報処理学会論文誌，Vol.32，No.3，pp.396-403（1991）
9) 五百川，宮崎：痕跡処理のための逐次型一般化 LR パーザ SGLR の拡張，言語処理学会第４回年次大会発表論文集，pp.314-317（1998）
10) 藪，藤石，宮崎：表現構造と話者の認識構造を抽出する日本語文パーザの試作，言語処理学会第３回年次大会発表論文集，pp.205-208（1997）
11) 高草木，宮崎：SGLR-plus による話者の認識構造を抽出する英語文パーザの試作，情報処理学会第５８回全国大会，No.1E-8（1999）
12) 池原，宮崎，白井，横尾，中岩，小倉，大山，林：日本語語彙大系 全５巻，岩波書店(1997)
13) 太田，前川，宮崎：規則用例融合型の日本語複合名詞構造解析法，言語処理学会第３回年次大会発表論文集，pp.313-316（1997）