日英機械翻訳のための新しい構文構造

New syntax structures for J-to-E machine translation

立命館大学名誉教授／有限会社サイバープロ代表 **池田　秀人**

PROFILE: ３０年以上データベース技術の研究を進めてきたが、2004 年ころから自然言語処理の研究を始め、機械翻訳、言語教育 e ラーニングシステム、特許文書作成支援システムなどの研究を行っている。

1 はじめに

特許文書の機械翻訳の品質を向上させるためには、①特許文書特有の構造をうまく利用することと、②翻訳を前提とした構文解析をすることが重要である。

特許文書特有の構造とは、「番号による参照」と、「括弧付挿入文」がその代表的なものである。参照番号は、特許文書の並列コーパスでは、同じ数字で表される。そのほとんどは名詞句であり、この同じ番号を持つ名詞句は基本的に対応している。また、挿入文は 1 つの文で高々 1 つにあることが多く、挿入部の対応が容易にできる。この 2 つの特長を使えば、対訳文対の句対応（Phrase Alignment）の精度が飛躍的に上がる。

また、翻訳を前提とした構文解析では、文の構文構造の切り出し方に特別な工夫ができる。対訳文対の構造は、名詞句が名詞句に訳されるわけではない。ほとんどの機械翻システムが同一構造、同一品詞に翻訳しようとしており、これが不自然な翻訳文を生み出す結果となっている。

我々は、フレーズ事例ベースのアプローチで特許文の翻訳に取り組んできたが、そこで、不成功翻訳の 1 つの重要な原因が、「異品詞フレーズ翻訳対」にあることからこれを克服するために、構文要素の変換を含めた構文構造を認識することが重要であることを痛感した。この論文では、この異品詞フレーズ翻訳対に対するその方法と結果を報告する。

2 異品詞フレーズ翻訳対

まず、次の翻訳例を見てもらおう。

原文:「*図 30 と図 32 は、ガイド５を１本だけセンターに敷設するか、2本にしてガイドローラ３の外側に敷設するかの違いであり、同じガイドローラ３の車両で対応できる。*」

翻訳文:「*The difference between FIGS. 30 and 32 lies in the difference between one guide 5 at the center and two guides 5 on the outer sides of the guide rollers 3, and a vehicle having the same arrangement of the guide rollers 3 can be used with both.*」

この翻訳対をそれぞれ独立に句分解すれば、表１のようになる。

この翻訳対のフレーズ対応を見てみよう。

この 2 つの構文構造の違いと、対応を考えてみると、次のような違いがわかる。

- 文フレームの構造の違い：日本語文の文フレームは、2 つの単文が、主語を共有しているのに対し、英文では、主語が異なる 2 つの文からなっている。意味構造を考えると、2 文を「and」で接続した複文であるという構造を持っているので、「2 文を and で接続した場合、後続文の主語は省略する」という規則が適用されているものと考えられる。

表１　事例の翻訳対をそれぞれ句分解したもの

日本文	英文
S0=_ は、@v 連用形 :、_。([N1],[P2],[P3]) N1=_ と _([N4],[N5]) N4= 図 _([N: 30]) N5= 図 _([N: 32]) P2=_ か、_ かの違いである ([P6],[P7]) P6=_ を _ 本だけセンターに敷設する ([N8],[N: 1]) N8= ガイド _([N: 5]) P7=_ 本にして _ の外側に敷設する ([N: 2],[N9]) N9= ガイドローラ _([N: 3]) P3= 同じ _ の _ で対応できる。([N10],[N9]) N9= 車両	S0=_ and _.([S1],[S2]) S1=_ lie in the difference between _ , and _ ([N3],[N4],[N5]) N3=the difference between FIGS. _ and _ (N:30],[N:32]) N4=one _ at the center([N6]) N6=guide([N:5]) N5=two @npl: on the outer sides of the _ ([N6],[N7]) N7=guide rollers ([N:3]) S2=a _ having the same arrangement of the _ can be used with both([N8],[N7]) N8=vehicle

表２　事例の翻訳対のフレーズ対応

日本文	英文
S0=_ は、@v 連用形 :、_。([N1],[P2],[P3])	S0=_ and _.([S1],[S2])
P2=_ か、_ かの違いである ([P6],[P7])	S1=_ lie in the difference between _ , and _ ([N3],[N4],[N5])
N1=_ と _([N4],[N5]) N4= 図 _([N:30]) N5= 図 _([N:32])	N3=the difference between FIGS. _ and _ (N:30],[N:32])
P6=_ を _ 本だけセンターに敷設する ([N8],[N: 1])	N4=one _ at the center([N6])
N8= ガイド _([N: 5])	N6=guide([N:5])
P7=_ 本にして _ の外側に敷設する ([N: 2],[N9])	N5=two @npl: on the outer sides of the _ ([N6],[N7])
N9= ガイドローラ _([N: 3])	N7=guide rollers ([N:3])
P3= 同じ _ の _ で対応できる。([N10],[N9])	S2=a _ having the same arrangement of the _ can be used with both([N8],[N7])
N9= 車両	N8=vehicle

- 次に先行単文（日本語文では、N1+P2, 英文の S1）では、日本語文では、2 つの動詞句（P6,P7）で表現されているのに対し、英文では、1 つの単文（S1）で表されており、日本語文の動詞（「違いがある」）が「difference」と名詞化されて「is 構文」に埋め込まれている。
- 日本語文の N1 には、「図 30 と図 32」という表現になっているが、この後に [の違い] が省略されており、対応する英語文の N3 では、この省略部「difference] が明示的に表現されており、かつ日本語文で冗長な「図」の繰り返しを英語文では避け、「FIGS 30 and 32」とまとめている。

翻訳時に行われるこのような品詞変換や冗長回避変換は、自然な文に翻訳するのに必要不可欠なものであるが、これが機械翻訳の品質の向上を妨げている原因の 1 つである。

3 構成要素の変換を含む構文構造

この問題を回避するため、表３のように変換操作を明示的に表現した構文構造を表３に示す。

表３　変換操作を追加した構文構造認識

日本文	英文
S0=@s 連動文化 ([S1],[S2])	S0=_ and _.([S1],[S2])
S1=_ は、_ か、_ かの違いである ([N3],[N4],[N5])	S1=_ lie in the difference between _ , and _ ([N3],[N4],[N5])
N3=_[との違い]([N6]) N6=_ と _([N7],[N8]) N7= 図 _([N: 3 0]) N8= 図 _([N: 3 2])	N3=the difference between _(N6]) N6=@n 共通名詞の統合化 (_ and _([N7],[N8])) N7=FIG_([N:30]) N8=FIG_([N:30])
N4=@s 名詞化 ([S9]) S9=_ を 1 本だけセンターに敷設する ([N10])	N4=@p 目的語による名詞化 ([P9]) P9=arranged one _ at the center([N10])
N8= ガイド _([N: 5])	N10=guide([N:5])
N5=@s 名詞化 ([P11]) P11= 2本にして [_ を]_ の外側に敷設する ([N8],[N12])	N5=@p 目的語による名詞化 ([P11]) P11=arranged two @npl: on the outer sides of the _([N12],[N8])
N12= ガイドローラ _([N: 3])	N12=guide rollers ([N:3])
S2=[_ は、] 同じ _ の _ で対応できる。 ([N3], [N13], [N12])	S2=a _ having the same arrangement of the _ can be used with both([N13],[N12])
N13= 車両	N13=vehicle

ここでは、次の変換関数が使われている。

- 連動文化関数（S0）: これは、２つの文の主語が共通の場合、主語を共通化し、後の文の主語を省略する変換をする
- 共通名詞の統合化関数（N6）: これは、２つの名詞句の先頭部が同じ名詞（句）になっている場合、共通部として使い、後の名詞句から共通部を省略する変換をする。上の例では、英語文のみこの関数が使われているため、日本語文のこの関数に対応する関数は無変換関数になっている。
- 文の名詞化関数（N4, N5）: この関数は、文を名詞句に変換するもので、上の例では、日本語文の場合、文がそのままの構造で上位の文（S1）使われているが、対応する英語文の場合、名詞化して上位の文（S1）に埋め込まれている。文の名詞化には、「主語」を名詞化する場合、「目的語」を名詞化する場合、「場所」「時間」、「方法」など中心動詞と格関係をもつ名詞句がその他の部分を修飾語としてもって名詞化される。上の英語文の例では、目的語 guide を被修飾語になるように名詞化されている。

この変換操作が文を構成するときどのように使われてきるかを明示的に表した構文構造が、「構文要素の変換を含めた構文構造」である。この構造を使って、文の句対応を抽出すると、文は文、節は節、述語は述語名詞句は名詞句に対応させることができると同時に、最上位の文は、どのような 変換を経て構成されるかが明確になる。英語らしい表現や日本語らしい表現は、文や節に現れることが多い。これは「言い回し」と曖昧に呼ばれているものに実体かもしれない。

4 おわりに

この論文では、フレーズ事例ベースの機械翻訳の対訳フレーズ辞書の作成のために、構文要素の変換を含めた構文解析の方法とその役割を示した。

このアプローチを使って、現在日英特許並列コーパス約 300 万文対から、対訳フレーズを抽出し、対訳フレーズ辞書を開発中である。語や基本述語レベルの対訳辞書は、既に多く実現されているが、文や節のフレーズの網羅的な辞書はまだ実現されていない。この研究では、その 1 歩手前まで来ている。この文や節の対訳プレーズには、字句的には似通っていない対応が多く含まれており、これが「〜らしさ」を表現していくのは興味深い。また、これは翻訳者のノウハウの 1 つであるが、これが辞書化されることで機械翻訳の質の向上が期待できる。