自然言語処理 112－8
（1996．3．14）

「の」を含む名詞句の日英翻訳に対する用例ベースアプローチ

飯盛 可織　佐川 雄二　大西 昇

名古屋大学大学院 工学研究科

〒464-01 名古屋市千種区不老町

名詞句と名詞句を「の」で結んだ「XのY」という名詞句は、日本語でよく現れる表現である。この名詞句では名詞同士が様々な意味関係を結ぶため意味解析が難しい。また英語への翻訳の場合、英語との対応が一対一ではないため翻訳規則を定式化することが難しい。用例を基に翻訳を行なう際に重要なのは、一致する用例が存在しない場合にどのような用例を当てはめて訳を行なうかである。本研究では、一般的に行なわれている名詞同士の意味距離を用いて用例を選択する方法の他に、名詞句と主部や述部との関係、並列関係などを手がかりにして翻訳パターンを決定する方法を提案する。

Example-based Japanese-English Translation for Japanese Noun Phrase "X 'no' Y"

Kaori Isagai, Yuji Sagawa, Noboru Ohnishi

School of Engineering, Nagoya University

Furo-cho,Chikusa-ku,Nagoya,460-01,Japan

The noun phrase patterns "X'no'Y", which consists of two noun phrases X and Y connected with particle "no", are frequently used in Japanese. Semantic structures of the noun phrases are difficult to analyze by a computer. When we try to translate Japanese into English, relations between the phrase and the corresponding English expressions are not simple. So it is difficult to formulate translation rules. The most important thing in translating with examples is to decide an apropriate example for transfer when there aren't the same examples. As usual, we select the examples, calculating the distance between the nouns. We propose a way to decide the examples according to the relations between the noun phrases and subject or predicate and so on.

1 はじめに

中間言語や中間表現を用いて機械翻訳を行なう場合、翻訳の対象となる言語を曖昧さのない構造で表す必要がある。しかし自然言語の性格として、意味を正確に記述することが難しい曖昧な部分は、必ず存在する。曖昧な部分が存在すると、その部分の構文構造や意味構造が決定できないため、正確な翻訳をすることができない。

日本語の場合、意味を正確に解釈あるいは記述することが難しい表現の一つに「名詞句＋の＋名詞句」という名詞句の構造がある。助詞「の」で結ばれる名詞句同士は、所有関係や場所など様々な意味関係を持ち得るため、全体の名詞句の意味構造を決定することが難しい。また、名詞句と対応する英語表現との関係が1対1でなく複雑であるため、機械翻訳を行なう際に翻訳規則を定式化することが非常に難しい。

このように意味解析の結果を正確に記述することが難しい場合、意味解析を行なうよりも、対訳の用例を用意しておいて置き換えを行う方式の翻訳の方が、正確な訳が得られる場合がある。しかし、準備する用例の数には限りがあるので、どの程度の数の用例を用意するのか、また翻訳対象と全く同じ用例がない場合に、どの用例を当てはめるか、といった問題がある。

本研究ではこういった問題を解決するため、一般的に用いられている名詞の意味距離の他に、名詞句の前後の表層的な情報を用いることで翻訳の精度を向上させることを試みた。

2 名詞句「XのY」

「の」は様々な意味関係を表すため、文法上の位置付けについても様々な説がある。本研究で扱う名詞句は、

> X、Yがともに名詞句、すなわちX、Yの主要部が名詞であるような名詞句「XのY」

また、名詞を

> 文法的に自立語で活用がなく「が」を伴って主語の文節を作れるもの

と定義する。

以下「XのY」は以上の定義を満たす名詞句を指すものとする。

島津らは「XのY」に対する訳の決定を試みているが、XとYを前置詞で結ぶ形、Yを前置詞として訳す場合、XまたはYを形容詞として訳す場合など、10のパターンに訳型を分けている [1]。一般的に「XのY」は前置詞"of"を用いて訳されると考えられているが、この分類では前置詞を用いて訳を行なうパターンのものが44.9%であり、"of"を用いるものはさらにこの中の一部である。このように、「の」を含む名詞句に対応する英語表現は非常に多様なパターンになる。

用例ベースの翻訳では、辞書が用例の形で登録されており、入力と辞書の用例の比較を行ない、最適な用例を用いて翻訳を行なう。訳のパターンが多ければ当然、可能な変換パターンの候補をどのように選ぶか、さらに最適なパターンをどのように選ぶかという問題が生じる。

「の」を含む名詞句の文法的な曖昧さや複雑さを考えると、構文レベルの情報のみで正しい訳を得ることは難しい。そこで、意味レベルの情報を用いることが必要になる。意味レベルの情報を扱うためにはそれぞれの単語の意味の記述を行なわなければならないが、意味の記述をどの程度行なえばよいかという基準は存在しない。さらに、実際には一つの単語の持つ意味が一つであるとは限らない場合もある。例えば「学校」という名詞だが、これは、場所、組織など複数の意味を持つ。これでは、単語の意味の記述を正確に行なったとしても、結局どの意味を選択するかが問題となる。そこで、文法的な情報だけでなく意味的な情報を含めて、最適のパターンを選ぶ手がかりをいくつか考える必要がある。

まず、X、Yの主要部となる名詞である。二つの名詞句のXの主要部が同じ場合は、Yの種類によって、そのXとYの関係は大いに変わり得る。例えば、前に述べた「学校」という例だが、「学校の授業」では、Xは組織としての学校を表しているが、「学校の隣」では、Xは学校の場所を表す。

3 機械翻訳の手がかり

ATR自動翻訳電話研究所によるATR対話データベースを対象とした [2]。これは、100万語の規模の対話データベースである。日英の対訳データが存在し、対訳のみでなく様々なデータが付与されている。付与されているデータは、会話ID、文ID、文節ID、単語IDなどの他に、品詞コード、活用形コード、文節対応ID、単語対応IDなどがある。対話の内容は、国際会議に関する問い合わせである。

今回は、単語テーブルに付与されているデータの中から、格助詞のコードを付与されている「の」の

みを対象とした。格助詞のコードを与えられている「の」について、その「の」を含む文節とその後の文節を取り出すことから、「Xの Y」という句を取り出す。さらに、文節対応IDを用いて、その文節と対応する英語の部分を取り出した。そして、本研究で対象とするX、Yがともに名詞句である名詞句「Xの Y」だけを取り出した。ここでは、その結果得られた名詞句1535例について、分析を行なった。

このデータのうち、354例は名詞句に対応する部分が存在しなかった。残りの1182例のうち、日英の対応が全く同じものを除くと、630例となる。しかし、特にこのデータは会話による通訳を通したものなので、文脈によってはきちんとした対応がとれていないものも多い。

このうち、Xの主部が指示代名詞のものが50例あるが、この場合、照応の問題を考えなければならないので今回は扱わない。その他には、固有名詞を含むものが45例、日付や時刻が関係しているものが53例、日付以外の数字が関係あるものが49例存在する。

その他の一般的なもの450例のうち、前後の文脈を考慮しなければ適切な訳が行なわれないものが26例あった。さらに、このコーパスは同時通訳を通した対話のものなので、意訳が多くて直接の対応をとることができないものも多く存在するが、こういったものは翻訳のベースとなる用例に用いることはできない。

以後、比較的文法的な拘束の強いと考えられる日付及び時間に関するもの、数詞を含むものと、曖昧な意味を持ちにくい固有名詞の場合を論じた後、その他の一般的な場合についての訳の決定法について述べる。

3.1 日付及び時間に関するもの

日付や時間に関する名詞句は、

i) 「9月の20日」など、YがXをさらに特定するようなもの

ii) 「5日の日」など、日時の強調を行なうもの

iii) 「10分間の質疑応答」など、あるイベントの期間を表すもの

iv) 「5日の開会式」など、イベントの日時を特定するもの

v) 「会議開催の1カ月前」など、イベントとの相対的な時間関係をあらわすもの

に分けられる。

i)の場合は、文法的に訳がはっきりと決定される。また、「1987年の6月の20日」、「1987年6月の20日」、「1987年の6月20日」は「1987年6月20日」と同じである。

実際には具体的な日時を表す数が変わっても翻訳パターンは同じなので、用例として辞書に登録する場合は、属性をもちいて登録すべきである。次に例を示す。

「9月の20日」:"the 20th of September"
→ ((month),(day)):(Y of X)
「5日の日」:"the 5th"
→ ((day), 日):(X)
「10分間の質疑応答」:"10 minute question and answer"
→ ((time) 間, 質疑応答):(X Y)

3.2 数に関するもの

X、Yのいずれかに数詞を含む名詞句について考える。

Xのみが数字を含む場合は、助数詞がついているものがほとんどであった。大きく分けて、Xの助数詞がYを数えているものと、Xが助数詞の数えているものを表している場合のどちらかとなる。前者の例としては、「12,000円の登録料金」、「3つのホテル」、後者の例としては、「3方の登録」、「2インチのスライド」などである。「3方の登録料金」は、「3方の人の登録料金」であり、「2インチのスライド」は「2インチの大きさのスライド」を表している。

この場合、助数詞が何を数えているかがわかる場合が多いので、その数えているものとYの名詞との関係でパターンが決定できることが多いと考えられる。

Yが数詞を含むものもやはり助数詞を持つ。例としては、「全額の10%」、「パンフレットの2ページ目」などである。

日付などの場合と同様、具体的な数値が変わっても訳のパターンが変化するわけではないので、辞書に用例として登録する際には次のようになる。

「12,00円の登録料」:"12,000 registration fee"
→ ((num) 円, 登録料):(X Y)

「3つのホテル」:"3 hotels"
→((num) つ, ホテル):(X Y)
「3方の登録」:"the registration for the 3 members"
→((num) 方, 登録):(Y for X)
「2インチのスライド」:"2 inch slides"
→((num) インチ, スライド):(X Y)
「全額の10%」:"10 percent"
→(全額,(num)%):(Y)
「パンフレットの2ページ目」:"the second page of your pamphlet"
→(パンフレット,(num) ページ目)
:(Y of X)

3.3 固有名詞に関するもの

X、Yのいずれかに固有名詞を含む名詞句について考える。

固有名詞は比較的その語が表す対象がはっきりしている場合が多い。データ中に現れた固有名詞は、地名、建築物名、組織名、人名であった。

例としては、次のようなものがある。

- 「東区のツインタワー」:
"Twin Tower in Higashi ward"
- 「大和銀行の大阪支店」:
"Daiwa Bank in Osaka"
- 「アメリカのワシントンディーシー」:
"Washington D.C., America"
- 「山形大学の坂本」:
"Sakamoto at the Yamagata University"

これらの場合も日付や数などと同じように、名詞の具体的な内容が変わっても訳のパターンが変化することはない。ただし、人名と地名では全く異なるものになる。最低限、地名、建築物名、組織名、人名を区別する必要がある。そこで、辞書に用例を登録する際は次のような形で行なう。

「東区のツインタワー」
→((noun_place),(noun_build))
:(Y in X)
「大和銀行の大阪支店」
→((noun_org),(noun_place) 支店)
:(X in Y)
「アメリカのワシントンディーシー」
→((noun_place),(noun_place))
:(Y, X)
「山形大学の坂本」
→((noun_org),(noun_man))
:(Y at X)

3.4 用例の種類

第3.1節、第3.2節および第3.3節で挙げた例は、それぞれで述べたように用例として辞書に登録する際は、"(num)" や"(noun_man)" など抽象化して表すことにした。以後は、一般的な名詞からなる名詞句と同様に扱う。

そこで今まで述べた例を含めて、用例辞書を構築するという観点から名詞句の性質について述べる。

まず、単純なパターンに置き換えができるものとできないものがある。単純にパターン化できないものは、サ変名詞が動名詞や不定詞の形になっていて、単純な形に置き換えられないものや、意訳になっているものなどである。それ以外のものは、X、Yがそれぞれの名詞句の直訳の形、X'のように" ' " のついたものなどが形容詞や前置詞として訳されているものとし、前置詞やそのX、Yの順序などで単純なパターンとして表す。

パターンの例:
「会議の概要」:the outline of the conference
→(会議, 概要):(Y of X)

パターン化したものを、別な名詞で置き換え可能かということと、文脈によって複数のパターンと置き換え可能かどうかということを考えると次のような場合が考えられる。

i) 慣用表現、イディオムなどの用例。「XのY」を一つの決まった表現に置き換えるもの。

例:「今のところ」
→(今, ところ):(at the moment)

ii) 一般的な用例。特別な制約なしにXやYの意味の類似度が高い用例を用いることができるもの。

例:
「開会のスピーチ」:"the opening speech"
→(開会, スピーチ):(X Y)

「開催の辞」:"the opening speech"
→(開催, 辞):(X Y)

iii) 汎用性はあるが、文脈によって用例を決定する必要があるもの。

例:
「日本人の方」:"Japanese"
→(日本人, 方):(X)
「日本人の方」:"Japanese speakers"
→(日本人, 方):(X Y')

ここで" Y' " は、"Y" に当たる部分に直訳ではないものが当てはまる場合である。

翻訳を行なうためには、入力された名詞句がどのパターンに当てはまるかを区別しなければならない。

i) のような慣用表現的なものは、「X の Y」全体で一つのまとまった意味を持つので、X と Y を別々に扱うことはできない。正しい訳を得るためには慣用表現専用の辞書が必要である。これは、池原らなどの従来の機械翻訳の手法と同様である [3]。

次に、一般用例を用いた場合のことを考える。古瀬らによる用例ベースを用いる翻訳では、入力と用例を比較して単文や句構造などそれぞれのレベルでのパターンマッチングや、単語の意味距離などを用いて、最も近いと判断される用例を用いて変換が行なわれる [4]。名詞句同士で適切な用例を選ぶ際には、入力と用例との名詞同士の意味の比較を行なっている。しかし iii) にあるように、一般用例による置換えだけでは正しく置換えが行なわれないことがある。そこで、単純な単語の意味距離計算だけでなく、文脈を考慮してパターンの決定や省略部分を補うような過程を考え、一般用例をそういった過程が必要なものとそうでないものに分けて考えた。

日本語特有の表現として形式名詞があるが、形式名詞は無視して良い場合もあるが、意味を前後の文脈から推定して訳語を与えなければならないような無視できない場合もあり、文脈を必要とすることが多い。

3.5 文脈を必要とする場合

まず、文脈を必要とする例を示す。

X が地名の場合、「" 地名 " の " 形式名詞 " 」のときは訳のパターンが「X」という形になる場合が多いが、例外として、次のものがある。

「京都の方(ほう)」
→"the geography in Kyoto"

これは、前後の関係から意味を詳しく限定できる場合である。この名詞句が現れた例文を次に示す。

> 私ちょっと京都の方は不案内なんですけれども、どの線、国鉄ですか、私鉄ですか、どれを使えば一番早く行けますでしょうか。
> I'm actually not very familiar about the geography in Kyoto.should I take private railway or J.N.R., which is faster?

「" 地名 " の方」という場合、その名詞句を目的格としてもつ述部によって、" 方 " の持つ意味が変わってくる。述部が「行く」などの動詞の場合は、その名詞句は地名の場所を表すものと考えられる。しかし、述部が「不慣れである」や「不案内である」などの場合は、英語ではきちんとその場所の地理であるということを表記するような場合でも、日本語では省略していることが多い。例の場合は、その後に交通手段のことについての問い合わせがあることからも、「京都の方」は意味的には「京都の地理」であることがわかる。

述部以外に用いることができる文脈の情報としては、次のようなものが挙げられる。

- 接続詞:「日本人の方」(日本人の方ならば、) →"Japanese"
- 主部:「企業の方」(300 名のほとんどが企業の方) →"from private company"
- 並列:「会議の参加者」(会議の参加者、聞かれる方々はどうでしょうか) →"the audience"

4 システム

第 3 章で挙げられた手がかりを元に、図 1 にあるようなシステムを提案する。まず用いる辞書について詳しく述べる。

4.1 辞書について

用いる辞書は、用例辞書、シソーラスと対訳辞書である。

用例辞書について

用例辞書は句構造レベルでの変換知識を記述したもので、対訳辞書は単語レベルの変換知識を記述した辞書である。用例辞書の方は、第 3.4 節で触れた通り、慣用表現的用例、文脈を必要とする用例、一般用例の 3 つのタイプの用例に分けて記述されている。

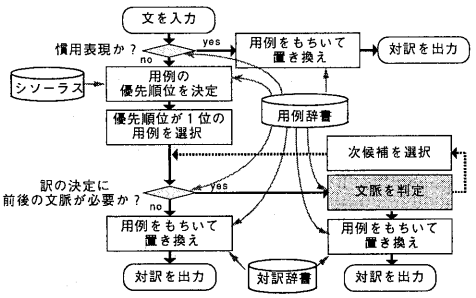

図 1: システムの構成

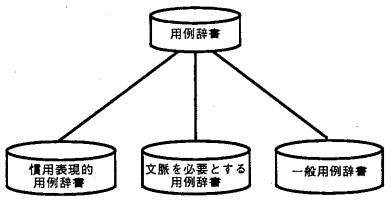

図 2: パターン辞書の構成

シソーラスについて

今回は、シソーラスとして分類語彙表を用いた[5]。次の例に示すように、意味範疇ごとにまとめてコードを割り振ったものである。意味的に近いカテゴリーのものは、近いコードが与えられるように考慮されている。

コード	見出し語	語
1.411	紙	紙 ペーパー パルプ
1.412	木・石・金	用材 コルク 板 石材
1.413	燃料・肥料	燃料 マッチ 石炭 硫安

表 1: シソーラスの例

単語辞書について

単語の辞書は日本語のそれぞれの名詞に対して、対応する英語表現が記述された辞書である。

4.2 処理の流れ

1. 慣用表現用辞書とのマッチング

入力は「Xの Y」を含む一文とするが、最初に計算を行なうのはX、Yの部分のみである。まず最初に、用例辞書の中の慣用表現用用例辞書を用いて、見出し語となっているX、Yと入力のX、Yについてパターンマッチングを行ない、同じものが見つかればその対訳パターンを出力結果とする。

2. 見出し語の意味距離計算による汎用パターンの決定

慣用表現辞書の見出し語と一致しない場合は、残りの汎用パターンの見出し語と入力のX、Yをシソーラスを用いて意味の類似度の計算を行ない、最も名詞同士の類似度が高いと判断されたパターンを用いて翻訳を行なう。

類似度の計算は、シソーラスを用いて行なう。入力とパターンのX、Yのシソーラスのコードを調べる。入力のXと汎用パターンのX_p、入力のYと汎用パターンのY_pのコードを比較する。入力とパターンの距離は次の式で計算することにする。

$$d((X,X_p),(Y,Y_p)) = |d(X,X_p)|*w_1+|d(Y,Y_p)|*w_2 \quad (1)$$

ここで、w_1、w_2は重みである。これらの式と条件によって計算を行なった結果、最も意味距離の小さい用例から順に対訳の候補として、次の処理へ進む。

3. 文脈情報を必要とする用例を用いる場合

辞書に記述されている文脈情報のタイプを元に解析を行ない、その結果を出力とする。文脈情報の手がかりが得られなかった場合は、次に意味距離の小さい用例を候補として前の処理へ戻る。

4. 一般用例を用いる場合

意味距離計算の結果選ばれた用例が一般用例の場合は、一般用例用辞書に登録してある用例に従って翻訳を行なう。まず最初にX_j、Y_jについて単語辞書をもとに単語の変換を行なう。変換して得られた単語X_e、Y_eをtarget patternのX、Yにそれぞれ代入することで、目的言語表現が得られる。

4.3 処理例

ここで、文脈を必要とする場合と必要としない場合の処理例を示す。

文脈を必要とする場合

入力:「大阪 (noun_place) の方」
(私は大阪の方が不慣れですが、)

- 慣用表現的用例辞書には登録されていない
- 類似度が最も高いものが2パターン存在した

用例
((noun_place), 方) : (X)
((noun_place), 方) : (geography in X)
(predicate:(不案内、不慣れ))

- 文脈を必要とする用例を先に候補として選択する
- この場合、述部に「不慣れ」が存在するのでこの用例による置き換えを行う

出力:"geography in Osaka"

文脈を必要としない場合

入力:「大阪 (noun_place) の方」
(新幹線で大阪の方へ行きます。)

- 文脈を必要とする用例を先に候補として選択するところまでは、文脈を必要とする場合と同じ
- 入力には述部に「不案内」、「不慣れ」といった語が現れないので、この用例は不適切である。
- 次に類似度の高い用例を選択する。この場合はXとYが同じ一般用例である。
- 用例による置き換えを行う

出力:"Osaka"

4.4 結果と考察

文脈を考慮することで、形式名詞などの意味を補って翻訳を行うことができるようになったものが、26例中16例存在した。英語表現に比べて曖昧な日本語表現から、英語表現への変換を行なうことができた。

用例ベースによる翻訳は、用例を収集したコーパスの内容に左右される。今回は対話のコーパスを用いており、対話の内容も国際会議に関する問い合わせに限られている。そこで、簡単なスクリプトなどを用いて、話者が質問者と事務局の職員であることや、話題が会議についてであることなどの状況の記述を行なうことで、さらに適切な翻訳を行なうことが考えられる。

5 おわりに

翻訳や意味解析が難しいとされている「の」を含む名詞句について、実際にある名詞句の分析を行ない、用例ベースによる日英翻訳の手法について考えた。特に、形式名詞などについてその意味を補うために、名詞句の述部や並列関係などを用いることを提案した。今後の課題としては次のようなことが挙げられる。用例の優先順位を選択する際に、今のところはX、Yを対等に扱っているが、用例を決定する際にどちらかが重要な位置を占めている可能性がある。このような場合、単語の意味距離の計算において、XとYにつける重みを偏らせることが有効となる。用例によって重み付けを変化させることによって、より適切な用例を選択することができる。また、第4.4でも述べたが、文脈を表現する記述としてスクリプトなどを用いることが考えられる。

謝辞 今回研究に用いたデータベースの構築携わったATR自動翻訳電話研究所の方々に感謝致します。また、本研究を進めるにあたり、有益なコメントをいただいた名古屋大学大学院工学研究科の山村毅講師、名古屋大学大型計算機センターの田中敏光助教授、および情報第7講座の皆様方に感謝致します。

参考文献

[1] 島津, 内藤, 野村: 助詞「の」が結ぶ名詞の意味関係のsubcategorization, 情報処理学会自然言語処理研究会, No.53-1,pp.1-8(1986).

[2] 江原, 井ノ上, 幸山, 長谷川, 庄山, 森本: ATR対話データベースの内容, ATR Technical Report,1990.

[3] 池原, 宮崎, 白井, 林: 言語における話者の認識と多段翻訳方式, 情報処理学会論文誌' Vol.28,No.12,pp.1269-1279(1987).

[4] 古瀬, 隅田, 飯田: 経験的知識を活用する変換主導型機械翻訳, 情報処理学会論文誌, Vol.35,No.3,pp.414-425(1994).

[5] 国立国語研究所: 分類語彙表, 秀英出版,1964.