

定型パターンを含む文の機械翻訳手法

加 藤 直 人[†]

外電を対象とした、英日機械翻訳システムの研究を進めている。外電ニュースは政治、経済、スポーツ等多岐の分野にわたっており、使われる表現も様々である。しかし、経済ニュースのように定型表現が多い分野も多い。定型表現によって、経済ニュースは文型に独特な特徴があり、訳語も独特的の表現が要求されるので、従来のような解析一変換一生成という機械翻訳では処理が困難である。むしろ、定型表現を積極的に利用して、特別な方法で翻訳処理するほうが精度の向上を期待できる。本論文では、定型表現を多く含む文（定型文）を精度よく翻訳処理する手法について述べる。本手法では、定型文を翻訳する際に、定型文翻訳データ（FST データ）が重要な役割を果たす。FST データは、定型文の英単語列を記述した文脈自由文法のルールと日本語テンプレートから成り、定型文とその対訳から自動的に作成される。その定型文は、含有率を計算することにより大規模コーパスから自動的に抽出される。本手法によって、外電経済ニュースを対象とした定型文翻訳実験を行った結果、翻訳率が従来の我々のシステムより約 10% 向上した。

Machine Translation of Sentences with Fixed Patterns

NAOTO KATOH[†]

This paper proposes a method for translation of sentences with fixed patterns (or fixed sentences). We are developing an English-to-Japanese machine translation system for the AP (Associated Press) wire service news stories. Some of news topics in the news stories, such as economy, sports and etc., are hard to be translated by conventional rule-based machine translation methods, because these topics have many idiomatic expressions or fixed patterns which are difficult to be processed by conventional syntactic analysis and/or word selection methods. In the proposed method, FST (Fixed Sentence Translation) data play a main role of translating fixed sentences. The FST data consist of a list of fixed English patterns and Japanese templates, which are built automatically from fixed English sentences and their corresponding Japanese translations. The fixed English sentences are extracted automatically from a large corpus. A series of experiments was conducted to evaluate the proposed method using economic news in the AP news stories. The translation accuracy is about 10% higher than our rule-based machine translation method.

1. はじめに

我々は、毎日大量に送られてくる英文外電を対象とした、英日機械翻訳システムの研究を進めている。外電ニュースは政治、経済、スポーツ等多岐の分野にわたっており、使われる表現も様々である。しかし、分野を限ってみると定型表現が多く使われる分野も多い。

特に、経済の話題を扱ったニュース（経済ニュース）は定型表現を含む文が多い。

経済ニュース文は

f 1) 英文は特殊な文型をしている文が多い。

f 2) 日本語に翻訳する際に、高品質の翻訳をするには経済独特の日本語訳が要求される。

という二つの大きな特徴を持つ。このような文を、従来の機械翻訳のような解析一変換一生成という過程を経て翻訳すると、それぞれの特徴は

p 1) 構文解析のための文法ルールの数が足りない。

（しかし、単純に増やすと構文的曖昧性が増えるという問題があらたに生まれる。）

p 2) 訳語選択が困難である。

という問題を生じる。これらの問題は解決困難な問題であり、実際に翻訳率も低い。（我々の従来の翻訳システムの場合、約 20% である。）

しかし、定型表現を多く持つという特徴は、経済ニュースは文の種類がそれほど多くないという特徴でもある。したがって、これらの定型表現を積極的に利用

[†] 日本放送協会 放送技術研究所 先端制作技術研究部
Program Production Technology Research Division,
Science and Technical Research Laboratories, NHK
(Japan Broadcasting Corporation)

して、定型表現で大部分被覆された文（定型文）のみを特別の方法で翻訳する方が精度の向上が期待できる。

このようなアイデアに基づき、定型文を精度良く機械翻訳する手法について述べる。本論文では、特に、本手法で重要な役割を果たす定型文翻訳データ（FST データ）の自動作成を中心に述べる。FST データは、英語の単語列を記述した文脈自由文法のルールと日本語テンプレートから成り、定型文とその対訳から自動的に作成される。その定型文は、含有率を定義することにより、大規模コーパスから自動的に抽出される。

最近、定型表現を積極的に利用した機械翻訳の一つに、用例に基づく手法が提案されている^{1)~3)}。古瀬ら^{4),5)}は、従来からのルールに基づく機械翻訳と定型表現の翻訳処理とを融合した翻訳手法を提案している。この方法は、定型表現等の用例をあらかじめ保持し、翻訳では入力と用例との類似度を計算するというベストマッチであるので、ある程度の定型さでも扱えるという優れた面を持つ半面、適切に類似度が計算できない限り、正しい翻訳が保証されないという欠点を持っている。同様なことは事例ベース翻訳やメモリベース翻訳などと呼ばれる推論を用いる方式すべてにあってはある。これに対し、本論文で述べる定型文の翻訳処理は、ルールに基づく機械翻訳とは独立に行う処理であり、この処理で翻訳できなかった文はあらためて従来の機械翻訳で翻訳する。このように独立した処理を行うことにより、古瀬らの類似度に頼る方法とは異なり、定型文翻訳で処理された文は翻訳率が 100% となる利点がある。

用例による機械翻訳にはそこで使う翻訳データを作成することが重要であるが、Kaji ら⁶⁾は、原文と対訳をそれぞれ構文解析し、その中に含まれる内容語の対応をとることによって、自動的に 2 言語間のテンプレートを作成する手法を提案している。本手法で使われる FST データも原文と対訳から自動的に作成されるものであるが、原文は構文解析するのではなく、定型表現を認定しその訳語を生成しているので、原文を部分的に構文解析するだけである。さらに対訳も構文解析はせず、原文中の定型表現に対応する訳語を対訳文中で照合するだけである。したがって、得られた翻訳データは Kaji らのほど応用範囲の広いものではないが、本作成手法には原文も対訳も完全な構文解析処理は必要ないという利点がある。

FST データを作成するにはその対象となる定型文を集めなければならない。定型文を定義し、大量のコーパスから定型文を直接抽出する手法はないが、定型

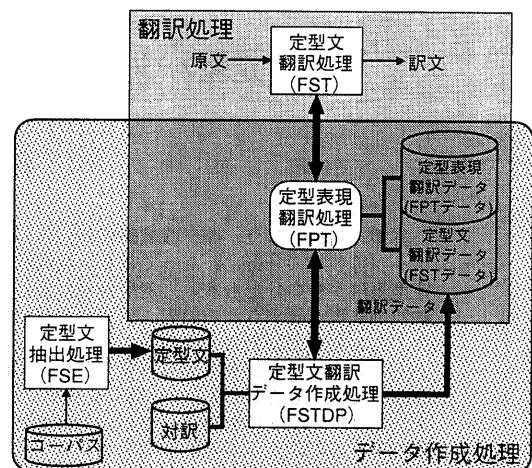


図 1 定型文処理システム概要
Fig. 1 Overview of processing fixed sentences.

表現を自動的に抽出する方法は様々提案されている。例えば、北ら⁷⁾は仕事量基準を用いて定型表現を自動抽出する手法を提案している。しかし、連続する語のみの抽出であり、二つの単語列間にギャップを持つ定型表現は扱っていない。Smadja ら⁸⁾は 2 単語間の距離を考慮することによりギャップを持つ場合でも定型表現を抽出する方法を提案しているが、ただ二つの単語間の距離、すなわち一つのギャップしか考慮していない。したがって、二つ以上のギャップを持つ定型表現を抽出することが困難である。本論文で述べる定型文の自動抽出手法は、二つ以上のギャップをも扱うことができる文単位の抽出手法である。

定型文機械翻訳手法は、次の四つの処理から構成されている（図 1）。

- s 1) 定型文翻訳 (FST)
- s 2) 定型表現翻訳 (FPT)
- s 3) 定型文翻訳データ自動作成 (FSTDTP)
- s 4) 定型文抽出 (FSE)

FST は、入力された英文が定型文であれば翻訳し出力する。FPT は、FST および FSTDTP に共通する処理系であり本手法の中心をなす。FPT は、定型表現の構成を表わしたルールとその構成要素となる語を登録した辞書を使って、入力された文から定型表現を抽出するとともに、その日本語訳、品詞、意味マーカを与える。FST では処理の際に定型文翻訳データ (FST データ) が必要となるが、FSTDTP は、定型文とその対訳から FST データを自動的に作成する。FSE は、その入力となる定型文をコーパスから自動的に抽出する。以下、2 章では定型文機械翻訳手法を構成する FST, FPT, FSTDTP, FSE のそれぞれについて述べる。3 章では、

本手法を外電コーパスに適用し、定型文を自動抽出した実験結果と、定型文が頻出する外電経済ニュースを翻訳対象に限定した、定型文機械翻訳実験の結果について述べる。

2. 定型文機械翻訳手法

出現頻度が高く、意味的なまとまりを持つ単語列のことを定型表現という。定型パターンとは出現頻度が高い単語列のことをいうが、意味的なまとまりをなさないものも含める。また、本論文で扱う定型文とは、2.5節で厳密に定義するが、定型表現や定型パターンで大部分被覆された文のことをいう。例えば、2.1節に示す例文1では、数量表現である“17.76 dollars per kilo”が定型表現，“Malaysian tin closed at”が定型パターン、例文1が定型文である。

外電ニュースの中でも経済ニュースは、定型表現である数量表現、日付表現等が変化するだけという文が多く、一般的のニュースに比べると定型文が多い。したがって、翻訳結果を得るだけであれば、従来のような機械翻訳は必要なく、単純に訳語の置き換えて翻訳できる場合が多い。本論文で述べる定型文の機械翻訳手法の基本的アイデアは、訳語の置き換えるである。

2.1 定型表現翻訳(FPT)

はじめに、定型文翻訳(FST)と定型文翻訳データ自動作成(FSTD)で中心的な役割をもつ定型表現翻訳(FPT)^{9)~11)}について述べる。

実用上、機械翻訳システムの精度向上には辞書の充実が大きく寄与するが、数量表現等は数字部分が変化するので辞書に登録しておくことはできない。我々の翻訳システムでは、形態素解析の後に設けられた処理FPTによって、ニュース文に頻出する数量表現、固有名詞、日付・時刻表現等の定型表現は一つの連語として品詞、訳語、意味マーカが決まり、構文解析に渡される(図2)。

例文1が処理される場合を例にとり、FPTを説明する。

[例文1]

“In Kuala Lumpur, Malaysian tin closed at 17.76 dollars per kilo, up 5 cents”

この文の形態素解析が終了した後、FPTは数量表現“17.76 dollars per kilo”

“5 cents”

を抽出して処理し、

日本語訳	品詞	意味マーカ
「1キロ 17.76 ドル」	名詞	数量表現
「5セント」	名詞	数量表現

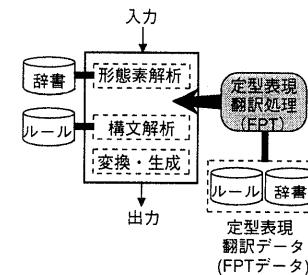


図2 FPTを含む機械翻訳システム概要
Fig.2 Overview of machine translation system with FPT.

英語定型表現	日本語テンプレート	FPT ルール
1: S --> unexp	「#1#」	
2: unexp --> unexp PER UNIT	「1 #2#1#」	
3: --> NUMEXP UNIT	「#1##2#」	
4: UNIT --> "dollar", "cents", "kilo", "yen", etc	「ドル」, 「セント」, 「キロ」, 「円」 ...	
5: PER --> "per", "a"	「」	
6: NUMEXP --> "1", "12", etc.	「1」, 「12」 ...	
7: CMA --> ","	「」	
8: UPDW --> "up", "down"	「アップ」, 「ダウン」	
9: CITY --> "Kuala Lumpur", "Tokyo", etc.	「クアラルンプール」, 「東京」 ...	

(ただし #i# はCFGルール右辺第i項に対応する日本語訳)を表す。また、「」は日本語訳がないことを表す。

図3 FPTデータの例
Fig.3 Example of FPT data for translating fixed patterns.

と日本語訳、品詞、意味マーカを決定する。

FPTは、具体的には、英語の定型表現の並びを文脈自由文法(CFG)で記述したルール(FPTルールと呼ぶ)と、FPTの処理対象となる語が登録されている辞書(FPT辞書と呼ぶ)とを使って処理を行う。ただし、これらのCFGルールや辞書は図2に示すような機械翻訳システムの形態素解析や構文解析で使われる文法ルールや辞書とは独立したものである。

FPTルールとFPT辞書登録語の例を図3に示す。ここで、“S”は開始記号，“unexp”は非終端記号，“UNIT”, “PER”等は前終端記号を表し、本論文では非終端記号と前終端記号を総称して記号と呼ぶ。また、日本語テンプレート中の#i# (iは数字)はFPTルール右辺第i項に対応する日本語訳を表す。例えば、3番目のルールでは、#2#はUNITの日本語訳(「ドル」, 「セント」, ...)である。以下では、このiをルール位置と呼ぶ。

図3中、1~3行目がFPTルールの例であり、4~9行目がFPT辞書に登録された語の例である。ただし、6行目は数字列が辞書に登録されているように表現しているが、実際には数字文字処理によって、数字部分の辞書項目が翻訳処理時に動的に作成される。また、

2行目の日本語テンプレート「1 # 3 ## 1 #」の最初の「1」は、例えば、「1 キロ 17.76 ドル」の「キロ」の前の「1」であり、固定的な日本語訳である。

特徴的なのは FPT ルールに日本語訳を生成するためのテンプレート（日本語テンプレート）が付随していることである。FPT では変換・生成処理は特別には設けず、日本語テンプレートを用いて日本語への変換・生成を行う。具体的にはボトムアップ CHART 法¹²⁾によって、英文を解析し、解析途中にルールが適用された際には、その記号に対応した日本語訳をテンプレートの変数に代入するという単純な操作によって日本語を生成する。このような日本語生成により、日本語テンプレートは CFG ルールごとに記述できるので、定型表現ごとの柔軟な日本語生成が可能となるという利点がある。例えば、“5 cents”では、“5”と“cents”は図 3 中の 6 行目と 4 行目の辞書登録語からそれぞれ訳が「5」、「セント」とまず得られる。3 行目のルールを適用して活性弧が NUMEXP, UNIT と埋められるのにしたがい、日本語テンプレートは #1#=5, #2#=セントと代入され、「5#2#」、「5 セント」と日本語生成が行われる。品詞や意味マーカも同様に CFG ルールに付随した規則で決まる。

FPT は、実際には数量表現以外にも “U. S. President Bill Clinton”のような役職付き人名等も処理し、FPT ルールの総数は約 150、FPT 辞書登録語数は約 4,000（そのうち 60% は人名・地名）である。このような FPT ルールは、後述する AP 電の連続語データを参考にして作成し、FPT 辞書に登録する語は、定型表現を構成する単語列に含まれる単語の中で出現頻度の高い語を選んだ。以下では、FPT ルールと FPT 辞書を合わせて FPT データと呼ぶ。

2.2 定型文翻訳 (FST)

定型文翻訳 (FST) は、基本的には FPT において、定型表現という概念を定型文まで拡張したものである。FST は、後述する FST データを追加して拡張された FPT データを使って、定型文を翻訳する。

FST は、英文の解析においては、定型表現を記述するときと同様に定型パターンの並びを文単位に記述した CFG ルールにより行われる。しかし、FPT とは異なり、英文を構成する定型パターンは意味的文法的なまとまりを持つものだけではない。また、日本語の生成は、変数化した部分を持つ文単位の日本語テンプレートにより行われる。

例文 1 を FST で翻訳するために追加した FST データの例を図 4 に示す。

図 4 では、FPT 辞書(図 3)に登録してある語“Kuala

英語解析CFGルール	日本語テンプレート
1: S --> PAT1 CITY CMA PAT2 UNTEXP CMA UPDW UNTEXP 「#2#でマレーシアのすずは、#8##7#の#5#でひけた」	
2: PAT1 --> "In"	「」
3: CITY --> "Kuala Lumpur"	「クアラルンプール」
4: CMA --> ","	「」
5: PAT2 --> "Malaysian tin closed at"	「」
6: UPDW --> "up"	「アップ」
7: untexp --> "17.76 dollars per kilo"	「1 キロ 17.76 ドル」
8: --> "5 cents"	「5 セント」
(untexpはFPTによる)	

図 4 FST によって例文 1 を翻訳するための FST データ
Fig. 4 FST data for translating Example 1 by FST.

Lumpur”や FPT で処理される数量表現 “17.17 dollars per kilo” 等を除いた、英単語列 “In”, “Malaysia tin closed at” に前終端記号 PAT1, PAT2 と、一つの単語列に一つの前終端記号をそれぞれ割り当てて FPT 辞書に追加する。これらの英単語列は意味的文法的なまとまりをなすとは限らない。そして、これらの記号を英文中の出現順に並べて、FPT ルールに追加する。

こうして、定型文を翻訳するために追加された部分の FPT 辞書、FPT ルールを特に FST 辞書、FST ルールと呼び、両者をまとめて定型文翻訳データ (FST データ) と呼ぶ。翻訳の際には、定型表現部分を翻訳する FPT データと、定型文全体を翻訳する FST データの両方を使って処理する。

FST データはサンプルされた定型文 (サンプル定型文) から自動的に作成されるが、詳細は 2.3 節で述べる。

FST データが追加されると、FST は定型表現の部分が異なる文でも翻訳できるようになる。例えば、図 4 のような FST データが FPT データに追加されており、図 3 のように FPT 辞書に “down” があらかじめ登録されていると、次の例文

[例文 2]

“In Tokyo, Malaysian tin closed at
1941 yen per kilo, down 19 yen”
は、

「東京でマレーシアのすずは、19 円
ダウンの 1 キロ 1941 円でひけた」
と翻訳できる。

“down” のようにあらかじめ FPT 辞書に登録しておく語は、コーパスから抽出した定型文の中から、出現頻度が高いものを選択すればよい。

定型文を定型パターン列で記述することは、もちろん正規文法でも可能である。しかし、我々は将来、句

単位の定型表現とルールに基づく翻訳を協調的に処理することをめざしており、その際に、一般的な文法ルールとの融合のしやすさを考慮してCFGで記述した。

2.3 定型文翻訳データ自動作成 (FSTDP)

FSTは、サンプル定型文のそれぞれに対して、FSTデータを必要とするが、人手によってこれらを大量に作るのは煩雑である。しかし、FSTルールは単なる定型パターンの並びであり、言語的手法を用いていないので、自動的に作成するのは比較的簡単である。

FSTDPは、サンプル定型文と人手で与えたその対訳とから、FSTデータを自動的に作成する。FSTDPの処理アルゴリズムを図5に示す。また、このアルゴリズム中のSTEP2で行う点数計算のアルゴリズムを図6に示す。

図5のアルゴリズムを説明する。STEP0では人手によって対訳を与える。STEP1では、日本語テンプレートで変数となる英単語列の候補を求める。候補となるのは、FPTで処理される単語列やFPT辞書に登録されている単語である。実際には、入力英文 $w_1 \cdots w_n$ (w_i は i 番目の単語, n は語数) に対して FPT の処理を行い、非活性弧として前終端記号や CFG ルールでまとめられた非終端記号とその日本語訳として求める。ただし、非活性弧には開始記号 S で構成される弧は含めない。

STEP2では STEP1 で得られた候補に点数付けを行う。位置 i から位置 j までの非活性弧の点数 $\text{path}(i, j)$ は、その日本語訳が対訳中にあるかないかによって、図6の手続きにより求める。ただし、単語 w_i と単語 w_{i+1} の間に番号 i を付け、この番号を位置と呼ぶ。また、文中の最初の単語の左側を位置 0、最後の単語の右側を位置 n とする。点数は 2 よりも大きい数（計算のしやすさから 3 にした）の指數関数で表わすことにより、単独の語よりも連続語が選択されるようにした^{*}。STEP3では最適な非活性弧の組を選択。非活性弧の範囲によって文は様々に区切ることができるが、FPTで処理された単語列が最も多く文中に含まれるように選択する。これは、後述する定型文抽出で行うのと同様に、ダイナミックプログラミング (DP) による。STEP3で、対訳の中で変数化される日本語訳の部分が決まる。STEP4では、STEP3で選択されなかった英単語列に対して、左から順に連続する英単語列を切り出し、それぞれに前終端記号を与える。そして、

* 点数を 2 の指數関数にすると 2 語連続の場合、これを一つの単語と見たときの点数は $2^2 = 4$ 、1 語ずつとみたときの点数も $2 + 2 = 4$ となり、点数だけでは前者を選べなくなってしまう。

STEP0

人手で対訳を与える。

STEP1

英文 $w_1 \cdots w_n$ をFPTで処理。

STEP2

図6により非活性弧の点数を計算。

STEP3

DPにより最適な非活性弧の組を選択。

STEP4

FSTで処理されなかった英単語列に対して適當な前終端記号を与え、CFGルールを作成。

STEP5

対訳中、STEP3で得られた日本語訳は CFG ルールのルール位置に置換し、日本語テンプレートを作成。

図5 FSTデータを自動作成するためのアルゴリズム
Fig.5 Algorithm for constructing automatically FST data.

```
for i := 0 to n-1 do
    for j := i+1 to n do
        if (位置 i と位置 j の間に非活性弧があり,
            その日本語訳が対訳中にあるか)
        then
            path(i, j) = 3j-i
        else
            path(i, j) = 0
```

図6 非活性弧の点数計算
Fig.6 Algorithm for calculating points between w_i and w_j .

この前終端記号と STEP3 で得られた記号を出現順に並べ、開始記号 S を左辺とする CFG ルールを作成する。STEP5 では対訳中で、STEP3 で得られた日本語訳を、STEP4 で作成した CFG ルールのルール位置に置き換えて日本語テンプレートを作る。

例文1のFSTデータが自動的に作成されるようすを使って、FSTDPを説明する。

STEP0では例文1に対して日本語訳、

「クアランプールでマレーシアのすずは、5 セン
トアップの 1 キロ 17.76 ドルでひけた」
を人手で与え、例文1とともにに入力とする。

STEP1では、例文1にFPTを実行し、処理が終了すると、CHARTの中に解析の途中結果として非活性弧が図7のように得られる。例えば、“17.76 dollars”や“17.76 dollars per kilo”は非活性弧となり、非終端記号はともにuntexpとなる。また，“,”は元々日本語訳がないので、対訳との照合は成功するものとする。

STEP2では、例えば、非活性弧として得られた単語列 “17.76 dollars per kilo”（語数4）の日本語訳「1 キロ 17.76 ドル」が対訳中にあるので、図6の手続きにより点数が 3^4 と計算される。

STEP3では、DPにより、文中に含まれる語が最大となるのは、図7の実線で表された非活性弧を選んだ

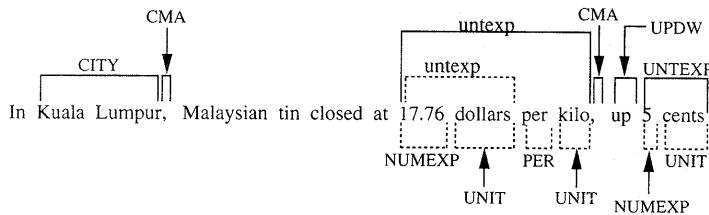


図7 FPTによる例文1の非活性弧
Fig. 7 Nonactive edges of Example 1 by FPT.

ときであり、最大値は、 $108 (=3^2+3+3^4+3+3+3^2)$ と求められる。

STEP 4では、STEP 3で処理されなかった英単語列には、それぞれに適当な前終端記号を自動的に作り(例えば、PAT 1, PAT 2, …というように),

PAT 1--> “In”

PAT 2--> “Malaysian tin closed at”

と割り当て、FST データに追加する。こうして、FPT 辞書に登録されていない単語列(それぞれの単語自身は低頻度である)が固定パターンとして扱われる。そして、得られた記号を出現順に右辺に並べて、左辺を開始記号 S とする CFG ルールで表すと入力英文のパターンの並びは、

S--> PAT 1 CITY CMA PAT 2

(ルール位置 1 2 3 4

UNTEXP	CMA	UPDW	UNTEXP
5	6	7	8)

と得られる。このとき、ルール位置も決まる。

STEP 5では、対訳中にある非活性弧の日本語訳「クアラルンプール」、「1キロ 17.76 ドル」、「アップ」、「5セント」をそのルール位置で決まる変数 "#2#", "#5#", "#7#", "#8#" に置き換える。すると、対応する日本語テンプレートは「#2#でマレーシアのすずは、#8##7#の#5#でひけた」と得られる。

この例からわかるように変数になる語はFPT 辞書に登録してあるものである。また、FST データを作成する際に、単純に数字のみを置き換えずに、数量表現を変数にしたのは次の理由による。数字のみの置き換えにすると、同じ数字が一つの英文中に2回出現した場合、その数字は対訳にも2回出現しているので、アルゴリズムのSTEP 2において対訳中で照合するとき、どちらの数字に対応するのかを決定できない。例えば、

[例文 3]

英文 “... Tuesday, off 0.74 points or 0.74 percent

from Monday's...”

対訳「…から 0.74 ポイント (0.74 パーセント) …」では、“0.74”が「0.74 ポイント」の「0.74」なのか、「0.74 パーセント」の「0.74」なのか、決定できない。しかし、数量表現とすることにより，“0.74 points”と“0.74 percent”の“0.74”がそれぞれ、対訳中「0.74 ポイント」、「0.74 パーセント」の「0.74」に対応することがわかる。

もちろん、対訳は数量表現で決め、FST データの中では数字のみの置き換えにするという方法もあるが、簡便さと他の表現の処理との整合性のために上記の方法をとった。また、同じ数字の数量表現が2回出現する場合もあるが、その場合には完全自動ではなく、人手によって対話的に選択する必要がある。

2.4 FST データ自動作成上の問題

今までの説明では、定型表現はFPTで得られた日本語訳と、人手で与えた対訳中の日本語訳が一致するものとしてきた。しかし、一般には一致しない場合もある。例えば、例文1の数量表現“17.76 dollars per kilo”は

T 1) 「17.76 ドル/キロ」

T 2) 「1キロにつき 17.76 ドル」

と翻訳することも可能である。この場合、自動作成ではこの数量表現全体が変数とはならず、単語単位の置換になってしまう。例えば、T 1)の場合には、

” 17.76 dollars per kilo ”

記号	untexp	PER	UNIT	「#1#/#3#」
----	--------	-----	------	-----------

(ルール位置 1 2 3)

となり、全体が一つのまとまり“untexp”で表せず、CFG ルールの右辺が長くなってしまう。この問題を避けるためには、FPT ルールを人手で作成する際に、考えられる日本語訳に対して異なったCFG ルールを作成する必要がある。例えば、図8のようにする。図8では、同じ英語表現に対して異なる日本語訳がある場合には、それぞれの日本語訳の数だけ英語解析ルールを分けた。そのため、ルール数が増えるという欠点はあるが、T 1), T 2)の場合とも数量表現全体をとらえた

1: S --> unexp	「#1#」
2: --> unexp1	「#1#」
3: --> unexp2	「#1#」
4: --> unexp3	「#1#」
5: unexp --> NUMEXP UNIT	「#1##2#」
6: unexp1 --> unexp PER UNIT	「#1###1#」
7: unexp2 --> unexp PER UNIT	「#1#/#1#」
8: unexp3 --> unexp PER UNIT	「#1#/##3#」

図 8 修正された FPT データ

Fig. 8 Modified FPT data for translating a variety of numerical expressions into Japanese.

英語パターンの並びを自動作成できるようになる。

定型表現に限らず、FPT 辞書に登録する語にも同様の問題が生じる。これもまた、異なる日本語訳ごとに英単語に異なった前終端記号を与えることによって対処できる。

2.5 定型文抽出 (FSE)

FSTDP では入力となる定型文が必要となる。定型文抽出 (FSE) は、大規模コーパスから自動的に定型文を抽出する。

FSE は、はじめに、あらかじめ数字や曜日等の単語の同一化処理を行った後、連続して出現する単語列(連続語)の頻度を求める。ある頻度以上出現した連続語のデータ(連続語データ)を用いて、これらの連続語が 1 文中に何%含まれているか、すなわち、

含有率 = 連続語の語数 / 総語数 × 100 (%) (1)
を各文に対して計算する。あるしきい値(P_0)を決め、 P_0 以上の文を定型文と定義する。コーパスの各文に対して含有率を計算し、 P_0 以上の文を定型文として抽出する。

例えば、次の例文

[例文 4]

- 1) The NYSE's composite index rose 0.39 to 196.61.
- 2) The NYSE's composite index edged up 0.33 to 186.51.

の含有率を計算することを考える。連続語データとしては、

“The NYSE's composite index rose 数字 to 数字”
(8 語連続)、

“The NYSE's composite index” (4 語連続),
“数字 to 数字” (3 語連続)

が得られているとする。それぞれの含有率は、

- 1) $8/8 \times 100 = 100(\%)$
 - 2) $(4+3)/9 \times 100 = 77.8(\%)$
- となる。

“rose”を含む文 1) は、文それ自体が連続語データとして求められていたので、従来の連続語の定型表現抽出手法でも得られるが、“edged up”は出現頻度が低いので、単語列間のギャップとなり文 2) は得られない。しかし、本手法では含有率を導入することにより、ギャップとなる単語列の前後の単語列が出現頻度が高いので、 $P_0=70$ とすれば文 2) も抽出できる*。また、このようなギャップが複数あっても、本定型文抽出手法では、 P_0 以上の含有率を持てば定型文として抽出できる。

さて、含有率を計算する際に、文中でどのように連続語をとるかによって、含有率が変わってくる。例えば 8 個の単語からなる文、“w₁w₂w₃w₄w₅w₆w₇w₈”(w_i は単語)の含有率を求める場合を考える。ただし、その中に含まれている単語部分列の中で、

“w₁w₂”, “w₁w₂w₃”, “w₂w₃w₄w₅w₆”,
“w₃w₄”, “w₇w₈”

が連続語データに含まれているものとする。含有率を計算する際に、連続語同士が重ならないように(例えば、“w₁w₂” と “w₁w₂w₃” が同時に選択されないように)どの連続語を選択するかによって、図 9 のような三つの場合が生じる。

含有率を計算する際には、一番多くの語数を含む区切り方が適切であると考えられるので、図 9 の場合は c) の 7 語を含む場合で含有率を計算したい。そこで次のようにして、一番多くの語数を含む区切り方を選ぶ。まず各位置間に点数を、その位置間に挟まれる単語列が連続語データに含まれていればその語数と、含まれていなければ 0 と定義する。次に以下に述べるように DP により、位置 0 から n までの点数の和の最大値と、そのときの含まれる単語列を求めることができる。

数式で表すと、位置 i から j ($i < j$) までの点数を

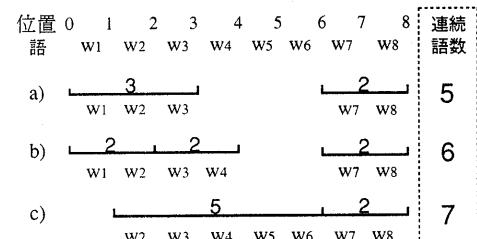
図 9 “w₁w₂w₃w₄w₅w₆w₇w₈” の語の区切り方

Fig. 9 Three cases in dividing Example
“w₁w₂w₃w₄w₅w₆w₇w₈”.

* 単語列間のギャップの存在を FSTDP で利用しているわけではないので、翻訳データの中では必ずしも変数となるわけではない。

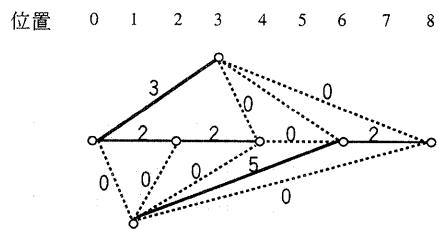


図 10 “ $w_1w_2w_3w_4w_5w_6w_7w_8$ ” 中の連続語で決まる点数
Fig. 10 Points of fixed patterns in Example
“ $w_1w_2w_3w_4w_5w_6w_7w_8$ ”.

point(i, j) とし,

$$\text{point}(i, j) = \begin{cases} j-i & w_{i+1} \cdots w_j \text{ が連続語データに} \\ & \text{入っているとき} \\ 0 & \text{入っていないとき} \end{cases} \quad (2)$$

$$\text{score}(j) = \max_{0 \leq k < j} (\text{score}(k) + \text{point}(k, j)) \quad (3)$$

で、最大語数 $\text{score}(n)$ を求めればよい。得られた値を一文中の連続語の語数とし含有率を計算する。図 9 の例では図 10 のように位置間の点数を決めることができるので、最大値は、1 文中に含まれる連続語の語数が 7 のときであり、含有率は

$$\text{含有率} = 7/8 \times 100 = 87.5\% \quad (4)$$

と計算される。

以上の定義と手法に基づいて、

- ・連続語データに含める連続語には、何語連続の語を使うか。
- ・どのくらいの出現頻度以上の連続語を使うか。
- ・含有率のしきい値 P_0 をいくらにするか

というパラメータをあらかじめ決定しておくことにより、大規模コーパスから定型文を抽出できる。

このようにして得られた定型文からサンプル定型文を選び（例えば、経済関係を翻訳対象とするならば、経済ニュースの文を選ぶというように）、FSTDP によって FST データを作成する。FSE と FSTDP とはまったく別の処理であるので、FSE で得られた連続語が FSTDP で必ずしも一つにまとめられるわけではない。

3. 実験

2 章では、定型文を機械翻訳するという立場から、FST, FSTDP, FSE の順で説明したが、実際には、この逆に処理する必要がある。したがって、実験結果は FSE, FSTDP, FST の順で述べる。

3.1 定型文抽出(FSE)実験

コーパスから FSE の処理実験を行った。コーパスには、AP 通信社 (the Associated Press) から配信され

るニュース、AP 電('89年4月～'91年3月の約2年間分、総文数約160万文)を使った。あらかじめ数字や曜日等の同一化処理を行った後、2～10語の連続語を求めて¹³⁾。ここで、同一化処理は次のように行った。

例えば、例文 1 の数字の同一化処理は、数字部分を数字[NUM]に変換して、

[例文 5]

“In Kuala Lumpur, Malaysian tin closed at [NUM] dollars per kilo, up [NUM] cents” とし、同様に連続語データも “[NUM] dollars per kilo”と変換して含有率を計算した。含有率がしきい値を越えていれば、元の例文 1 を出力する。

実験の結果、延べ約2万1千文（数字の同一化処理をすると、異なり文は約8千文であった）の定型文を抽出した。ただし、パラメータとしては、3～6の連続語を使い、その連続語の出現頻度は10回以上とし、またディスクの容量の都合で含有率は80%以上とした。

抽出された定型文の一部の例を付録に示す。実際の結果でもこの例同様、経済ニュース文が多かった。

3.2 定型文翻訳実験

3.1節の結果より定型文には経済ニュースが多いことがわかった。AP 電は一日350ほどのニュースが入電し、経済ニュースはそのうち50ほどである。各ニュースにはタイトルがついており、ニュースの特徴を表している。そこで、タイトルにより経済ニュースのみを選択し、定型文翻訳実験を行った。

FST データは、3.1節で得られた定型文から、経済ニュースに関する異なり文上位388文（平均19.4語/文）を人手で抽出し、人手で日本語訳を与え、FSTDP により作成した。得られた FST 辞書の語数は413語であり、英文1文あたりのパターンの数(S-->…で始まる右辺の項数)の平均は7.6項であったので、英単語は平均約3語ごとに前終端記号としてまとめられていることになる。

ほとんどのサンプル定型文が数量表現を含んでいたが、

[例文 6]

“Gold prices were mixed” 「金価格は小動きだった。」

のように変数を全く含まない文もあった。この場合には、FST データは、

S-->PAT 132 「#1#」

PAT 132--> “Gold prices were mixed”

「金価格は小動きだった。」

と日本語テンプレートに変数項を含まない、文全体として作成された。

次に得られた FST データを使って翻訳実験を行った。翻訳対象文にはランダムに選択した 2 日分の AP 電の経済ニュース、'91年 3月 7日(193 文)と 4月 25 日(167 文)分を使用した。前者は、定型文抽出時に利用したコーパスに含まれており、後者は含まれていない。

その結果、被覆率(=定型文翻訳処理で翻訳された文数/経済ニュースの全文数)は

3月 7 日 : $60/193 \times 100 = 31.1\%$

4月 25 日 : $53/167 \times 100 = 31.7\%$

となり、経済ニュース文の約 30%を本手法で処理することができた。翻訳された文はすべて正確に翻訳されていたので、適合率(=正しく翻訳された文数/FST によって処理された文数)は 100%である。我々の従来の機械翻訳システム(トランスファ方式)でこの経済ニュース文を翻訳したところ、被覆率は 100%であったが、適合率は約 20%であった。したがって、本手法の翻訳率(=被覆率 × 合成率)は約 30%であったのに対して、従来法では約 20%であったので、本手法だけでも約 10%向上した。本手法で翻訳できなかった文は、従来法で翻訳できる文もあるので、FST と従来法を縦続して組み合わせた翻訳システムの翻訳率は一層向上する。実際、FST の後に従来の機械翻訳システムを組み込み、FST で翻訳できなかった文は従来の機械翻訳で処理した、外電経済ニュース英日機械翻訳システム^{14),15)}を作成した。このシステムでは、FSTDP の入力にはならなかったがある程度定型パターンの含有率が高い経済ニュース文を集め、人手で文法ルールと辞書をこれらの文にチューニングし、FST で翻訳処理されなかった文はこの文法ルールと辞書で翻訳すること^{16)~19)}により、経済ニュースに関する文全体の翻訳率を 20%から約 70%に飛躍的に向上させることができた。

このような縦続した機械翻訳システムを構築した場合、翻訳速度が問題となる。しかし、我々の従来のシステムでは翻訳速度が 0.73 秒/語であったのに対して、FST の翻訳速度(実験で使った 360 文を処理した速度)は、0.08 秒/語と約 9 倍速い。したがって、トータルの機械翻訳システムでも、速度はあまり遅くならない。ただし、翻訳実験にはワークステーション SUN 3(約 3 MIPS)を使用した。

4. おわりに

FST, FPT, FSTDP, FSE の四つの処理から構成される定型文機械翻訳手法について述べた。また、本手法により、AP 電から定型文を抽出する実験を行い、

経済ニュース文にしほった定型文翻訳実験を行った。その結果、翻訳率が約 10%向上した。被覆率は 30%とまだ低いが、これは FST データを作成するときに用いる定型文の数を増やすことによって向上させることができ可能であり、それに伴って翻訳率も向上することが期待できる。

今後は、定型文抽出に関してはパラメータをいろいろ変え、抽出される文の比較検討を行いたい。また、定型文の機械翻訳処理をスポーツニュース、一般のニュースへ適用することも検討したい。

今回は文単位に定型表現を扱ったが、将来は句単位の翻訳にも拡張し、その処理で得られた結果と機械翻訳中の一般の文法ルール(CFG ルールで記述されている)を協調的に利用することをめざしている。

これらの課題を考察し、外電ニュース全体の翻訳率向上をめざす。

謝辞 本研究を進めるにあたって適切なご指導をいただいた二宮佑一部長(現、研究主幹)、ならびに相沢輝昭研究主幹、江原暉将主任研究員をはじめとする自動翻訳グループの方々に深く感謝する。また、本論文を書く上で浦谷則好主任研究員には適切な助言をいただいた。心から感謝する。

参考文献

- 1) Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, Elithorn, A. and Banerji, R.(eds.), pp. 173-180, North-Holland (1984).
- 2) 佐藤理史: MBT 2: 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) Sumita, E. and Iida, H.: Example-Based Transfer of Japanese Adnominal Particles into English, *IEICE Trans. Inf. & Syst.*, Vol. E-75-D, No. 4, pp. 585-594 (1992).
- 4) 古瀬 蔵, 飯田 仁: 変換と解析の協調的処理による翻訳手法—変換主導型翻訳手法—, 情報処理学会自然言語処理研究会報告, NL-87-4, pp. 27-34 (1992).
- 5) Furuse, O. and Iida, H.: Cooperation between Transfer and Analysis in Example-based Framework, *Proc. of Coling-92*, pp. 645-651 (1992).
- 6) Kaji, H. and Morimoto, Y.: Learning Translation Templates from Bilingual Text, *Proc. of Coling-92*, pp. 672-678 (1992).
- 7) 北 研二, 小倉健太郎, 森本 邸, 矢野米雄: 仕事量基準を用いたコーパスからの定型表現の自動抽

- 出、情報処理学会論文誌、Vol. 34, No. 9, pp. 1937-1943 (1993).
- 8) Smadja, F. A. and McKeown, K. R. : Automatically Extracting and Representing Collocations for Language Generation, *Proc. of ACL-90*, pp. 252-259 (1990).
- 9) 加藤直人, 浦谷則好, 相沢輝昭, 中瀬純夫: 英日機械翻訳における有名詞処理, 第40回情報処理学会全国大会論文集, 2F-2, pp. 421-422 (1990).
- 10) Katoh, N., Uratani, N. and Aizawa, T. : Processing Proper Nouns in Machine Translation for English News, *Proc. of International Conference on Current Issues in Computational Linguistics*, pp. 431-439 (1991).
- 11) 中瀬純夫, 鄭 敏: 英日機械翻訳における局所解析について, 情報処理学会自然言語処理研究会報告, NL-78-19, pp. 145-152 (1990).
- 12) 野村浩郷: 自然言語処理の基礎技術, p. 290, 電子情報通信学会, 東京(1988).
- 13) 浦谷則好, 加藤直人, 相沢輝昭: AP電経済ニュースからの定型パターンの抽出, 第42回情報処理学会全国大会論文集, 6E-4, pp. 178-179 (1991).
- 14) 加藤直人, 鎌田雅子, 相沢輝昭: 外電経済ニュースの英日機械翻訳, 第45回情報処理学会全国大会論文集, 2E-4, pp. 95-96 (1992).
- 15) 相沢輝昭, 鎌田雅子, 浦谷則好: 外電経済ニュースの英日機械翻訳: 新しいアプローチ, 電子情報通信学会言語理解とコミュニケーション研究会報告, NLC-91-45, pp. 65-72 (1991).
- 16) 相沢輝昭, 浦谷則好, 鎌田雅子: 外電経済ニュースの英文の定型パターン, 第43回情報処理学会全国大会論文集, 2H-1, pp. 187-188 (1991).
- 17) 相沢輝昭, 加藤直人, 鎌田雅子: 外電経済ニュース用英日機械翻訳システム, 電子情報通信学会言語理解とコミュニケーション研究会報告, NLC-92-39, pp. 61-65 (1992).
- 18) 相沢輝昭, 鎌田雅子: AP電経済ニュースの英語解析用文法, 第45回情報処理学会全国大会論文集, 3E-6, pp. 117-118 (1992).
- 19) Aizawa, T., Katoh, N. and Kamata, M. : Tuning of a Machine Translation System to Wire-Service Economic News, *Proc. of PACLING-93*, pp. 304-308 (1993).

付 錄

定型文の例（含有率；定型文）

100;Gold prices were mixed

100;He did not elaborate.

100;No injuries were reported.

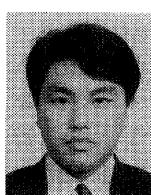
100;The U.S. dollar opened at 159.97 yen on the Tokyo foreign exchange market Monday, up from last Friday's close of 157.65 yen.
..... (中略)

97.0;The average price for strict low middling 1 1-16 inch spot cotton declined 99 points to 78.64 cents a pound Wednesday for the seven markets, according to the New York Cotton Exchange.
..... (中略)

85.2;Philippine peso banknotes Friday at 20.50-21.00 pesos (dealer buying-dealer selling) per U.S. dollar at the close, unchanged from a day earlier.
..... (後略)

(平成6年4月22日受付)

(平成7年6月12日採録)



加藤 直人 (正会員)

1962年生。1986年早稲田大学理工学部電気工学科卒業。1988年同大大学院理工学研究科博士前期課程修了。同年日本放送協会入局。放送技術研究所にて機械翻訳の研究に従事。1994年(株)ATR音声翻訳通信研究所に出向。音声言語処理、対話処理の研究に従事。電子情報通信学会会員。