# 市況速報文を対象とする用例利用型日英機械翻訳

5R－5

足立 貴行　高橋 大和　内野 一　古瀬 蔵　白井 諭
NTT コミュニケーション科学研究所

## 1. はじめに

ルールベース翻訳は要素合成的な翻訳のため，入力文に特殊な表現や語句の省略がある場合，高品質な翻訳ができないという問題点がある．一方，用例翻訳では事前に大量の対訳用例を準備しておくことで，様々な入力文に対して高品質な訳文を生成できる．対訳パターンを用いる用例翻訳の手法[1][2]が提案されているが，人手で大量の対訳パターンを作成するため，人的コストや時間がかかるという問題点がある．

本稿では，文対応付けされた大量の対訳用例を用い，その文をそのまま使用することを基本として高品質な訳文を生成する用例利用型日英機械翻訳[4][5]について述べる．本方式で有効な適用対象は，対訳用例を大量に集めることが可能で，ルールベース翻訳では高品質な翻訳結果を得ることが困難な文を多く含んでいる分野であり，現時点では市況速報文を対象として考えている．

## 2. 対訳コーパスの作成

本研究では大量の対訳用例を利用するために対訳用例の人手加工は行わず，文対応の情報がある対訳コーパスをそのまま利用することを前提としている．しかし，実際には人手で対応付けされた２言語のコーパスだけでは大量の対訳用例を収集するのは困難である．そこで，我々は同じ内容について記述された２言語のコーパスから自動的に文対応付けを行う研究[3]も進めている．現在は人手で対応付けしたものを利用しているが，将来は自動文対応付け機能を組み込み，大量の対訳用例が利用可能で高品質な訳文を生成する翻訳システムを実現する．

## 3. 市況速報文の特徴

市況速報文は継続的に日本語と英語の記事が配信されており，自動文対応技術が確立すれば大量の対訳用例を利用できる．現在，市況速報文として59386ペアの対訳用例を用いている．これらの市況速報文には以下のような特徴があり，用例利用型日英機械翻訳に有効な翻訳対象であると考えられる．

(1) 数詞や企業名などの固有名詞を多く含む
数詞を含む文は和文用例が30912文，英文用例が27068文存在．
(一般化しやすく，差分対象にしやすい)
- 「９時現在、前日比２７銭円安・ドル高の１ドル＝１０７円２２—２４銭。」
- 「京セラ、ＴＤＫ、アイワが上げ、三井不、菱地所が堅調。」
- 「日銀は５１００億円の資金不足に対し６０００億円の資金供給を通知。」

(2) 分野に特有な表現がある
(対訳英文に対訳が現れていれば翻訳可能)
- 「円高・ドル安」
  449文に存在．
- 「大証」
  2868文に存在．

(3) 語句(動詞)の省略がある
(省略に対する語句が対訳英文にあれば翻訳可能)
- 図２の入力文
- 「#D 時現在、前日比#D 銭円高・ドル安の#D ドル＝#D 円#D—#D 銭。」(#D は数詞の一般化，以下同様)
  20文存在．この20文の対訳英文に現れている動詞は，「traded」(7文)，「changed hands」(5文)，「was traded」(2文)．

(4) 定型的で類似した文を多く含む
(典型的な対訳用例を選択可能)
- 「#D ドル＝#D 円#D—#D 銭」を含む文
  159文存在．文の種類は異なり25種．

## 4. 用例利用型日英機械翻訳の処理

用例利用型日英機械翻訳(システム名：EUREKA)は，対象分野の文対応付けされた対訳コーパスを用いて，図１の手順で翻訳処理を行う．

以下，プロトタイプの各処理について述べる．

---

An Example-based Japanese-to-English Machine Translation for Short Market Reports
Takayuki Adachi, Yamato Takahashi, Hajime Uchino, Osamu Furuse and Satoshi Shirai
NTT Communication Science Laboratories

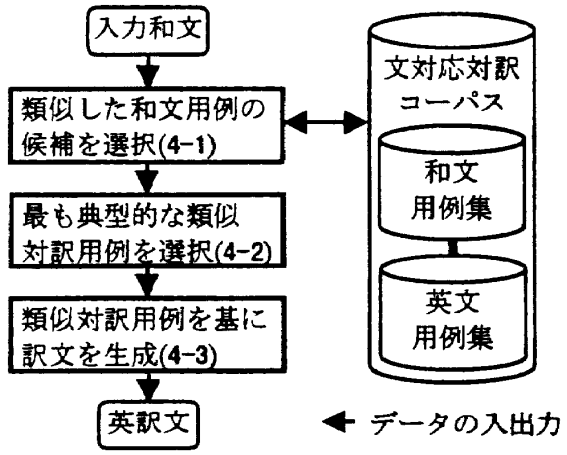


図１：用例利用型日英機械翻訳の処理手順

### 4-1. 類似和文用例の候補選択

大量の対訳用例から入力和文に類似する和文用例を高速に選ぶため，まず，和文用例の絞り込みを行う．事前に和文用例集に対し文字単位の N-gram 表現と文 ID を記録しておき，入力和文に含まれている N-gram 表現を調べ，その表現を持つ和文用例を候補和文用例として選択する．

### 4-2. 最も典型的な類似対訳用例選択

4-1 で得られた候補和文用例を用いて，(1)入力和文と候補和文用例の類似度，および候補和文用例の対訳である(2)候補英文用例の語による順位を考慮し，最も典型的な類似対訳用例を選ぶ．

(1) 入力和文と候補和文用例の類似度

文字列(形態素)ごとに区切り，候補和文用例の文字列を入力和文の並びにバブルソートしたときのスワップ数を基に計算する[4]．

(2) 候補英文用例の語による順位

候補英文用例の間で共通に現れる語(但し，冠詞，助動詞，前置詞を除く)の頻度が最大のものを典型的な語とし[4]，候補英文用例ごとに典型的な語の種類を多く含む文から順位付けする．

### 4-3. 類似対訳用例に基づく訳文生成

4-2 で得られた最も典型的な類似対訳用例の英文用例を模倣して入力和文の訳文を生成する．入力和文と類似和文用例の差分箇所，および類似和文用例の差分箇所に対応する類似英文用例の箇所を調べ，類似英文用例の差分対応箇所を入力和文の差分箇所の英訳で置換する．最後に，活用形などの細かい調整を行う．現在，市況速報文特有の表現である数詞や企業名などに着目して差分置換を試みている．

図２に本方式による市況速報文の翻訳例を示す．

●入力和文

Ij: 寄り付きの成り行き注文は、買い１０５万株、売り１０１万株。

●候補和文用例と候補英文用例
(入力和文の類似順，数詞は一般化して類似度計算，下線は候補英文用例の間で共通に現れる語の頻度が最大のもの)

cj1: 寄り付きの成り行き注文は、買い２０７万株、売り１０９万株。
ce1: Market buy orders at session opening stood at 2.07 million shares against 1.09 million sell orders.

cj2: 寄り付きの成り行き注文は、買いが２９３万株、売りは９９万株。
ce2: Buy orders at market opening came to 2.93 million shares compared with 990,000 sell orders.

cj3: 後場寄り付きの商いも１００枚強と薄い。
ce3: Trading was thin, with June drawing mere volume of 100 lots in the session.

●入力和文の英訳(cj1-ce1 をもとに訳出)

Oe: Market buy orders at session opening stood at 1.05 million shares against 1.01 million sell orders.

図２：用例利用型日英機械翻訳における市況速報文の翻訳例

## 5. おわりに

市況速報文を対象とする用例利用型日英機械翻訳について述べた．今後，プロトタイプ[5]を基に検討，改良を加え，ルールベース翻訳と用例利用型翻訳の長所を生かしたハイブリッド翻訳システムを構築する予定である．また，市況速報文以外の分野についても翻訳対象を拡大して研究を進めていく．

## 参考文献

[1]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.

[2]渡辺，武田：パターンベース翻訳システム：PalmTree，第 55 回情報処理学会全国大会講演論文集(2)，pp.80-81, 1997.

[3]白井，松尾，瀬下，藤波，池原：新聞記事日英対訳コーパスの構築 (3) —記事の特徴分析と文の対応関係の検討—，平成７年度電気関係学会九州支部連合会大会，p.857, 1996.

[4]S. Shirai, F. Bond and Y. Takahashi: A Hybrid Rule and Example-based Method for Machine Translation, NLPRS-97, pp.49-54, 1997.

[5]高橋，白井，立花，西垣，池原：用例利用型日英機械翻訳の基本設計，言語処理学会第３回年次大会，pp.145-148, 1997.