市況速報記事を対象とした日英翻訳システムの構成

5J－5

白井 諭*1 松島英之*2 井上浩子*2 松尾三津恵*2 矢部孝幸*2 内野 一*1

*1NTTコミュニケーション科学研究所 *2NTTアドバンステクノロジ(株)

1 はじめに

市況速報記事には，専門語や略語のほか，分野特有の表現が多用されるため，一般の機械翻訳を適用するのは難しい。しかし，類似の表現が多用されることから，それを考慮した専用の翻訳システムを構築して翻訳精度の向上を図ることが考えられる。

英語の経済ニュースに対して，原文の定型性に応じ翻訳テンプレート，経済専用文法，一般文法を切り替えて翻訳することにより，70%近い翻訳率が達成されている[相沢 96]。このような高精度を達成するには，翻訳対象にできるだけ特化することが必要であり，翻訳対象や記述言語ごとに課題となる項目が異なっている場合が考えられる。

本稿では，日本語の市況速報記事を対象に，見出しに対する翻訳処理，本文のうちの定型的表現に対するテンプレート型翻訳処理，ルール型翻訳処理を組み合わせた日英翻訳システムの構成を試みた。見出し翻訳では記事本文を参照することにより日英のスタイルの違いに対応する。また，補足表現に対処したり，ルール型翻訳の弱点を補強したりするため，前処理と後処理を導入した。

2 市況速報記事の英訳上の課題

本稿では，市況速報記事14週分（1995年6月～9月）のうち，東証外国部（1日1記事），大証（1日4記事），東証CB（1日3記事）に基づいてシステム構成を検討した。これらの記事は日本経済新聞社のテレコンデータベースから取り出し，[高橋 97]の方法により日英記事を対応付けた。英語は設計や評価の参考にした。記事の例を図1に示す。

3種類の記事とも，見出しの後，市場の総括が1文，概況や背景が1～3文，個別銘柄の様子が2～3文書かれている。

◇東証外国部・大引け
【NQN】ニューヨーク株の大幅高を映し堅調。売買高は概算20万株。売買の成立した39銘柄（値付き率48.1%）のうち、値上がり18、値下がり3、変わらず1、比較できず17だった。アップル、モトローラが上げ、IBMが年初来高値に顔合わせした。ボーイングは年初来高値更新。半面、グラクソWL、バークレイズ、CSHが下げた。
Tokyo Foreign Stocks Cls: Up on rally in N.Y.
(NQN) Foreign stocks ended higher Friday in line with the overnight run-up on Wall Street. Turnover was estimated at 200,000 shares. Among 39 issues changing hands, 18 increased, three declined and one issue settled flat. No comparison was available for 17 stocks. Boeing marked a year's high. IBM matched its year's high. Other gainers included Apple Computer and Motorola. In contrast, Glaxo Wellcome, Barclays and CS Holding lost ground.

◇東証CB大引け・5年ぶり大商いで続伸
【NQN】大幅高で14日続伸。株高から電機関連を中心に株価連動銘柄が買われたほか、債券高を材料に利回り銘柄も物色された。売買高は概算2000億円と90年5月16日の2300億円以来、約5年2カ月ぶりの大商い。ただ、利食い売りも目立ち、値上がり銘柄数395に対し値下がりも157とけっこうあった。住友精化(1)、富士通(9)(10)、NEC(8)(9)が高かった。一方、東ガス(3)、関西電(3)は軟調。
Tokyo CBs Cls: Up in active trading
(NQN) Convertible bonds registered their 14th consecutive day of gains Friday on trading centered around electric issues which track their underlying stocks and speculation in high-yielders. Turnover was a whopping 200 billion yen, the first time that trading volume has reached this level in 62 months. Despite the fact that gainers outnumbered decliners by 395 to 157, profit-taking resulted in losses for a considerable number of issues. The QUICK CB Index closed the day 1.55 points higher at 481.10. Sumitomo Seika Chemicals (No. 1), Fujitsu (Nos. 9 & 10) and NEC (Nos. 8 & 9) rose. Meanwhile, Tokyo Gas (No. 3) and Kansai Electric Power (No. 3) were weaker.

図1 日英の市況速報記事の例（1995年7月7日）

見出しは，英語には必ず総括の記述があるが，日本語には欠けることがある（図1の東証外国部）。本文の第1文から手がかりを得ることを考える。

市場の総括は，類似の表現が頻出するので，テンプレート化が期待される。しかし，英語記事には日付に相当する曜日が記述されているため，それを挿入する処理を設ける必要がある。

概況や背景は，文が長めで，因果関係が書かれることが多い。長文分割により精度向上を狙う。

個別銘柄の様子は極めて定型的であり，テンプレート化が期待される。

そのほか，括弧書きに対する対策が不可欠である。テンプレート化については[白井 97]で報告する。

3 翻訳システムの構成

前節の市況速報記事の特徴を踏まえ，図2のような翻訳システムの構成を考えた。各翻訳エンジンは

A Japanese-to-English Machine Translation System for Stock Market Reports

Satoshi SHIRAI*1, Hideyuki MATSUSHIMA*2, Hiroko INOUE*2, Mitsue MATSUO*2, Takayuki YABE*2 and Hajime UCHINO*1

*1NTT Communication Science Laboratories and *2NTT Advanced Technology Corporation

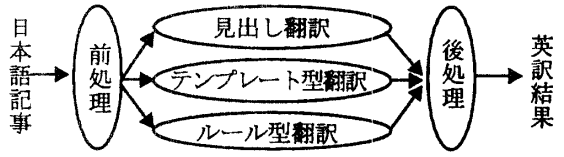

図２　市況速報記事翻訳システムの構成

並列に走行するが，自己評価の高い翻訳結果が一式揃った段階で翻訳結果を編集するようにした。

3.1 前処理

データベースから取り出した記事の文字コードを変換し，各翻訳エンジンに渡すデータを編集する。見出し翻訳処理には見出しと本文の第１文，それ以外の翻訳処理には本文を渡す。また，括弧表現を分類し[白井 96c]，翻訳対象であれば別の文にするとともに，埋め戻すための情報を後処理に引き継ぐ。

3.2 見出し翻訳処理

見出しは，分野（例えば「東証CB」），場（「大引け」），背景（「3 日続伸。」では「3 日」を背景とする），総括（「反落」），コメント（横線以降，欠落多し）から構成される。図１の東証外国部のように総括がなければ本文第１文から総括に該当するキーワードを取り出す。その結果，例えば96 年 9 月の市況速報記事 3 分野 80 記事の見出しのうち，77 件(94%)では英語記事と一致または同等の結果が得られた。失敗した３件は，コメント部分に対応するなどにより正解が得られる見込みである。

表１　見出し翻訳ルール数

種別	ﾙｰﾙ数	例（英訳）
分野	16	東証 CB(Tokyo CBs)
場	6	寄り付き,寄付(Opg), 前引け(Mng-cls)
背景	17	<数>日(for N straight days)
総括	98	反落(drop), 続伸(continue rising)
合計	137	

3.3 テンプレート型翻訳処理

形態素解析を行なった後，ルールと照合して変数部分を決定し，単一単語は対訳辞書引きで，複数単語はルール型翻訳で変数を英訳する[白井 97]。この方式は高速で，翻訳正解率もほぼ100%である。

3.4 ルール型翻訳処理

ベースとなる処理はALT-J/E[池原 91]である。対象記事の特殊性を考え，[相沢 96]のように専用文法と一般文法の２段階とはせず，専門分野対応のルールと辞書を使用した[白井 96a,b]。また，概況や背景は長文分割[白井 96d]により精度向上を図った。

3.5 後処理

各翻訳エンジンによる翻訳結果に基づいて，記事全体の翻訳結果を編集する。記事本文に対するテンプレート型翻訳とルール型翻訳の採否は次により決定した。テンプレート型の訳文は素点を 10 点とし，複数単語の訳出にルール型を使用した回数に応じ 1 点ずつ減点する。ルール型の訳文は素点を６点とし，未訳出は３点ずつ減点する。記事本文の後半の文はテンプレート型が適合しやすいため，多くの記事でルール型の翻訳完了前に翻訳結果が編集される。

この後，前処理で切り出した括弧表現を埋め戻すことにより，全体の体裁を整える。最後に，曜日記述の追加（図１の第１文末の Monday）や，冗長語の削除（例えば，大証では，見出しに Osaka とあるので，本文中の Osaka は不要）を行なう。

4 実験結果

翻訳結果は文単位で４段階評価し上位２段階を合格とした。記事単位では，全文合格なら合格とした。96 年９月の市況速報記事３分野 80 記事では，文単位の合格は 74%（◎61%,○13%）であるが，記事単位の合格は 23%（◎5%,○18%）にとどまる。しかし，東証外国部に限ると文単位で 90%（◎88%,○2%），記事単位で 70%（◎40%,○30%）に達する。分野によっては実用の見込みがある。

5 おわりに

本稿では市況速報記事を対象とした日英翻訳システムの構成を述べた。今後は，新たな翻訳エンジンの導入による概況や背景の文の翻訳率の向上を検討する予定である。

参考文献

[相沢 96] 相沢,加藤,鎌田：外電経済ニュースの英日機械翻訳，情報処理学会論文誌, Vol.37, No.6, pp.1041-1048
[池原 91] S. Ikehara, S. Shirai, A. Yokoo & H. Nakaiwa: Toward an MT system without pre-editing --effects of new method in ALT-J/E--, In Proc. of *MT SUMMIT '91*, pp.101-106
[白井 96a] 白井,井上,井田倉,池原,横尾：専門分野対応の日英機械翻訳用構文意味辞書の構築，言語処理学会第２回年次大会, A1-4, pp.13-16
[白井 96b] 白井,阿部,矢部,久保,池原,横尾：新聞記事日本文における書き替え対象表現の分布，言語処理学会第２回年次大会, A2-3, pp.37-40
[白井 96c] 白井,矢部,松尾,西垣,大山：新聞記事文における括弧書き表現の分析とその処理について，情報処理学会第 53 回全国大会, 2L-9, Vol.2, pp.31-32
[白井 96d] 白井,瀬下,木村,横尾,池原：従属節の階層構造に基づく日本語長文の自動分割とその効果，情報処理学会第 53 回全国大会, 4L-8, Vol.2, pp.67-68
[白井 97] 白井,細野,野沢,木村,阿部,内野：市況速報記事に対するテンプレート型日英翻訳の効果，情報処理学会第 55 回全国大会, 5J-6, Vol.2
[高橋 97] 高橋,白井,大山,渡邊,上田：日英新聞記事の記事対応コーパス自動作成，言語処理学会第 3 回年次大会, D1-4, pp.127-130