4G-1

電子メールを利用した機械翻訳システム

黒岩　眞吾　　松本　一則　　榊　博史
国際電信電話株式会社　上福岡研究所

1．はじめに

我々は、文単位の英日翻訳を行なう研究ベースの機械翻訳システムＫＡＴＥを開発した。このシステムを実際に翻訳支援システムとして用いる場合に必要となるツールを検討すると共に、ネットワーク上で電子メールを利用し機械翻訳サービス（英文メールを日本語に翻訳して返送する）を行なう場合の機能の分散化について検討を行った。

機械翻訳システムＫＡＴＥでは入力する英文に対し、(1) 入力される文の文区切り情報と、 (2) 英文の大文字小文字情報を翻訳時の情報として用いている。このため、ユーザーは前処理として、文区切り情報を付加すると共に、大文字のみで書かれている英文を大文字小文字を用いた文に変換する、といったプリエディット作業を行なう必要がある。これに対し本システムでは、粗訳でもよいから手間をかけずに結果が欲しいというユーザーを対象に、この二つの処理の自動化を試みた。

一方で、機械翻訳を利用する手法として、ネットワーク上の電子メール機能の利用を検討した。この場合、翻訳する文章に依存している機能は、ローカルに持たせる必要があるとの観点から、上記プリエディット機能およびユーザ辞書、未知語検索機能をローカルマシン上に置く手法を試みた。

2．電子メールを利用した機械翻訳サービス

電子メールを利用して機械翻訳システムを利用する利点としては、ネットワーク上のほとんどのユーザーが利用可能であること、計算機の負荷が低い時間に翻訳を実行できること、等があげられる。

機械翻訳において、ユーザー辞書が非常にに重要であることは周知の事実である。ＶＡＮなどで行なわれている翻訳サービスは、契約を結んだ特定のユーザーにのみサービスを提供することから、ユーザー辞書を各契約者ごとに、機械翻訳システム上に持たせている。しかし、ある程度不特定多数のユーザーを対象にした場合には、このアプローチは困難である。逆にすべてのユーザーが同一の巨大なユーザー辞書を利用する方式では、単語のユーザーごとの訳し分け、といったユーザー辞書本来の役割が薄れてしまう。そこで本システムでは、ユーザー辞書を各ユーザーがそれぞれのローカルマシン上に置き、ユーザー辞書による情報を英文テキストに付加してメールとする手法を用いた。

3．システム構成

3.1 翻訳の流れ

翻訳の流れを図１に示す。翻訳すべき原稿は、まず、スペリング誤りおよび未知語に関する情報を得るために「未知語検出」機構に入る。次に「文区切り抽出」機構によって文区切り記号を付加する。さらに、テレックスにより得られた原稿やニュースの表題のように大文字のみで書かれている原文に対しては「大文字小文字変換機構」により大文字および小文字を用いた文に変換する。そして、「ユーザー辞書引き」機構で、入力文をスキャンしてユーザ辞書に登録されている語を探索し、発見した語に関する辞書情報をテキストデータとして付加しメ

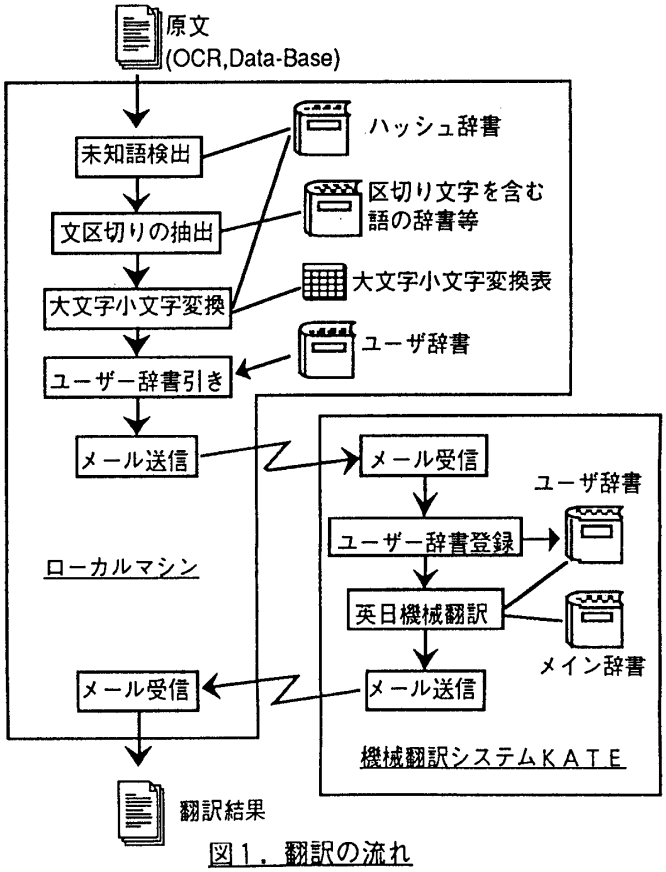

図１．翻訳の流れ

The machine translation system used under the electronic mail system.
Shingo Kuroiwa, Kazunori Matsumoto, Hiroshi Sakaki
KDD Kamifukuoka R.&D. Labs.

ールを作成する。作成したメールは機械翻訳システムにUNIXのメールシステムを通じて送信する。機械翻訳システムでは、受け取ったメールから辞書情報と本文を分離し、辞書情報をユーザー辞書登録した後、本文を機械翻訳し返送する。また、翻訳の終わった段階でユーザー辞書に登録した語を削除する。

3.2 文区切りの抽出

英文は一般には、.(period)、!(exclamaton mark)、?(queston mark) で文が切れると考えられるが、period を含む単語や、: (colon)、; (semicolon) 等に注意しなくてはならない。またデータベースから読み込んだテキスト等については、注意しないとタイトルと最初の文で一文としてしまう可能性がある。

本システムでは、以下のことを考慮し文区切りをプログラムで自動的に行なっている。

(1) 無条件で文区切りとなる記号
(2) 前後の単語情報により文区切となる記号
(3) 1行の長さ、及び次の行の先頭の単語情報

単語情報としては、単語長、大文字小文字情報およびperiodを含む語の辞書等、ユーザー独自の辞書情報を用いる。(1)の例としては空行、(2)の例としては period、また(3)はタイトルに対処するために必要な情報である。これらの情報に基づいた処理手順（例えば (2)の文区切り記号について、記号に続く単語の先頭が小文字ならば文は切れない等のルールの集合）はユーザーごとに設定可能とした。

3.3 大文字小文字変換

大文字のみで構成されている英文は、人間にとっても読みにくいものである。機械翻訳システムＫＡＴＥでは大文字小文字情報を積極的に利用して、形態素処理、及び文法解析を行なっているため大文字のみで構成されている文章を解析することは困難である。そこで本システムでは辞書エントリーのみを用いて、大文字小文字変換を行なう手法により変換プログラムを作成した。具体的には大文字のみで構成される語を辞書引きし、その結果（エントリーの有無）をもとに、表１の形の大文字小文字変換表にしたがって変換する。例えばCHINAという単語を辞書引きするとChina及びchinaのエントリーがあるので、表に従い(011) Chinaと変換する。この変換表と辞書の内容はユーザーごとに設定可能とした。（例えば As (砒素) は化学関係以外のテキストを扱う場合は辞書から除いておくことが望ましい）

辞書登録			
FOO	Foo	foo	出力
0	0	0	FOO
0	0	1	foo
0	1	0	Foo
0	1	1	Foo
1	0	0	FOO
1	0	1	foo
1	1	0	FOO
1	1	1	foo

表１．大文字小文字変換表

3.4 ユーザー辞書の分散化

ユーザー辞書をユーザーごとにローカルマシン上に持たせ、翻訳させたいテキストをあらかじめユーザ辞書引きし、ユーザ辞書に載っている単語・熟語情報をテキストに付加しメールとする。（図２）

機械翻訳ホストマシン上では受けとったメールを定められた時間にバッチとして機械翻訳する。この時、メール中の辞書情報の部分はあらかじめ切り出し、機械翻訳システムのユーザー辞書登録機構を用いて、辞書登録する。(図１)　そして、そのメールの機械翻訳を終えた段階で削除する。

```
To:honyaku@kate
Subject:tree
--------
.TS
Bell Atlantic's paging service also will be used as
system transport.
  ・・・
  ・・・
.TE
.DS
Bell_Atlantic ベル・アトランティック社 (10101100)
  ・・・
  ・・・
.DE
```

図２．翻訳のためのメール

4．結び

機械翻訳システムKATEをネットワーク上で利用する手法について述べた。今回のシステムは、実用というよりは主に翻訳結果の評価を行なってもらうために構成し、大勢の方から翻訳結果に関する評価を頂くことができた。

ユーザ辞書の分散化は、現在すべての情報をメールに付加しているが、訳語選択のみを付加する方法や、送られてきた辞書情報を蓄積する手法についても、今後検討していきたい。また、文区切り抽出や大文字小文字変換などを簡単にルールで記述できるような、文字列を扱うためのルール記述言語についても現在検討を進めている。

[参考文献]

[1] T.TANAKA : "THE PRESENT SITUATION AND PROSPECTS FOR MT-VAN SERVICE", IFTT'89, pp.38-44 (1989)

[2] 長谷部　他："ネットワーク接続される翻訳システムの辞書環境", 情報処理学会 第38回全国大会 (1989)

[3] 武田　他："ホスト/ パソコン機能分散型機械翻訳システム(1), (2)", 情報処理学会 第37回全国大会 (1988)

[4] H.Sakaki et.al : "A Parsing Method of Natural Language by Filtering