インターネット翻訳サービスユーザからの辞書データ収集

中山圭介 熊野 明

(株) 東芝 研究開発センター

〒210-8582 川崎市幸区小向東芝町1

{keisuke,kmn}@eel.rdc.toshiba.co.jp

1 はじめに

インターネットの普及により、多くのユーザが英語を中心とする外国語に触れる機会が増大し、その理解のために、機械翻訳のニーズが増大している。しかし、機械翻訳の品質は発展途上であり、精度向上には、新しい単語を含む辞書知識が不可欠である。

我々は、機械翻訳の最も基本的な知識である辞書知識を、実ユーザから直接収集する仕組みをインターネット上の翻訳サービス上で構築した。翻訳サービス MT Ave[1] は1997年7月8日からソフトウェア販売サイト SoftPark を通じてインターネット上で公開した。10月20日まで無料試験運用を行ったのでその結果について報告する。

2 システムの特徴

ユーザは Web 上のページで英語の原文を入力し、入力された原文は CGI を用いて翻訳サーバ側に送信される。サーバ側では英日翻訳を行ない、結果を電子メールでユーザに返送する。システムの構成を図1に示す。

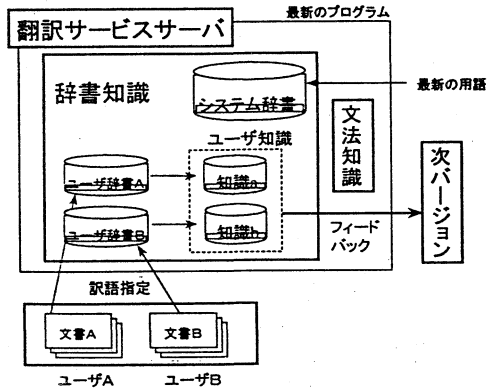

図 1: システム構成

2.1 訳語指定 (辞書登録)

ユーザは翻訳依頼の際、英語とその日本語の訳語をペアにして訳語指定を行うことができる。この機能は、従来の翻訳サービス [2] でも行われているが、本システムでは一度ユーザが指定した訳語はシステム側で自動的に蓄積され、以後の翻訳依頼でユーザが再利用できる。すなわちユーザは繰り返し翻訳サービスを利用する過程で、半自動的に個人用ユーザ辞書がサーバ側に構築される。[1]

2.2 用語抽出

ユーザの個人用辞書データベースの構築をサポートするために、用語抽出機能をシステムに組み込んだ。用語抽出機能は翻訳依頼英文を翻訳する際に、訳語指定すれば効果がありそうなもの (未知語等) を抽出してユーザに提示する。

3 翻訳量と訳語指定データ数

1997年7月8日の運用開始から10月18日までの3ヶ月間で約3,000人のユーザから約13,000件の翻訳依頼が得られた。この間の総翻訳量は訳文で約2千万文字におよび、サービスとしての注目度は高かった。

また、訳語指定 (辞書改良のための元データ) として、ユーザから5,860の生データが得られた。ユーザは翻訳依頼を行なう際、専門分野辞書を1つ選択して翻訳依頼を行なうことができる。訳語指定データ (ユーザから収集する翻訳知識) は、ユーザが翻訳の際に用いた専門分野辞書によって分類されてサーバ側に蓄積される。すなわち、ユーザがよりよい翻訳結果を得るために利用する専門分野辞書を1つ選択することにより、ユーザから収集する翻訳知識 (訳語指定データ) を半自動的に分野別に分類する。

[1] ユーザの辞書データを我々が利用できるよう、本サービスの利用条件にそのための一項を含め、サービス利用ユーザには必ずこれを承諾していただいた。

全体として5,860語の訳語指定データが得られたが、その分野別の内数は表1の通りである。

表1: 分野別訳語指定データの数

情報	1,629語
インターネット	950語
電気・電子	752語
化学	483語
機械	458語
政治経済	264語
分野指定なし	1,556語
全体[2]	5,860語

また、一ユーザあたりの訳語指定語数の分布を表2に示す。

表2: ユーザ別の訳語指定数

100語以上	7人
50語～99語	20人
20語～49語	39人
10語～19語	55人
5語～9語	118人
2語～4語	232人
1語	218人

4 訳語指定の自動分野分類

分野別の訳語分類に関しては、インターネットを通じた翻訳サービスのため、予想通り情報分野、インターネット分野の訳語指定データが多かった。また、どの専門分野の辞書をユーザが利用するかで、訳語指定データを自動分類する機能は多くの場合正確であり、本来情報分野に属すべき訳語指定が政治経済の分野に分類されていたり、化学の分野に属すべき訳語指定がインターネット分野に分類されている等の例は非常に少なかった。このため、分野分類が行なわれている訳語指定データに関しては、翻訳システムの専門辞書にフィードバックするために使用できそうである。逆に、ユーザが分野指定をしなかった(すなわち専門辞書を使用せずに翻訳した場合)ものに関しては、様々な分野のデータが入り混じっており、翻訳システムの辞書にフィードバックするためには、かなり分析が必要である。

また、現在専門分野辞書を6つしか用意していないので、ユーザが無理にこの6つの分野に分類してしまう傾向が見られた。例えば化学の分野での訳語指定には、生物関係の用語や、医学関係の用語もかなり含まれていた。また機械の分野には元来は物理学関係の分野に続する用語も含まれていた。また政治経済の分野には、銀行、金融関係の用語と、南北問題等国際関係の用語が両方含まれていた。例を以下に示す。

化学分野での訳語指定の例
```
cornea = 角膜
cultured tissue = 培養組織
epidemiology = 伝染病学
```

機械分野での訳語指定の例
```
activity = 放射能
aerodynamics = 気体力学
```

政治経済分野での訳語指定の例
```
checking account = 当座預金
infant mortality = 乳児死亡率
```

これらのお互いに似た関係にある分野の訳語指定をより詳細に分類し、翻訳システムの辞書強化により有効に利用するためには、予め用意しておくユーザが選択できる専門辞書の分野の数をもっと増やし、より狭い分野への分類が可能になるようにすることが今後必要である。専門辞書の数を増やすことは、現在分野未分類となっている訳語指定データに関しても、分野分類される機会を増やすことにより、より効率的な翻訳知識の収集を図れる。

上記分野分類のうち、情報分野とインターネット分野はかなり近い分野であり、ユーザが分野選択にとまどうことも予想されたが、収集された訳語指定データを見ると、インターネット分野では「MOUNTAIN VIEW = マウンテンビュー」などの訳語指定データが得られ、情報分野では「device driver = デバイス・ドライバ」等の訳語指定データが得られ、ユーザによる分野選択は、翻訳知識収集の観点からみて、かなり的確に選択されている。

[2]複数の分野で同じ訳語指定を行なっている場合があるので各分野の合計は全体とは一致しない。

インターネット分野での訳語指定の例

Cool Site = クールサイト
data encryption = データ暗号化
domain name = ドメインネーム
frame = フレーム
gateway = ゲートウェイ
Java platform = Java プラットフォーム
JavaScript = ジャバスクリプト
push technology = プッシュ技術

情報分野での訳語指定の例

BIOS = BIOS
GUI application = GUI アプリケーション
instance= インスタンス
bus reset state = バス・リセット状態
daemon = デーモン
device driver = デバイス・ドライバ
method = メソッド
registry = レジストリ

このことからも今後はより細かい専門用語辞書の分類を用意して、より詳細な分野情報が得られるようにすることは有用であると思われる。

5 訳語指定データの傾向

収集された訳語指定データの全体的な傾向は以下の通りである。

5.1 大文字で始まる単語についての訳語指定

文中に出て来る大文字で始まる単語の割合はそれほど多くないにもかかわらず、登録された単語の約 1/3 が大文字で始まる単語に対する訳語指定であった。

これはまず第一に固有名詞に対する訳語指定が多いことが理由である。固有名詞は特定のユーザの文書に固有のものが多く、それらについては翻訳システムでは未知語等となる場合が多いので、ユーザが固有名詞を登録することにより、大文字で始まる語の訳語指定が多くなった。

固有名詞の訳語指定の例

Damon Hill = デーモン ヒル
MOUNTAIN VIEW = マウンテンビュー
Hoover = フーバ大統領
Rochester region = ロチェスター地域

また第 2 には例えば

ITU = ITU
FDIC = 連邦保険預金機構
PPP = PPP
CPC = 臨床実践委員会

などの頭字語に対する訳語指定が多いこともその理由である。頭字語も固有名詞と同様に翻訳システムで未知語となる場合が多いので、それに対してユーザが訳語指定をする例が多かった。

5.2 原語をそのままのスペルで訳語指定

例えば

ANSI = ANSI
MIDI = MIDI
ping = ping
basic = basic

などの訳語指定が多数見られた。"basic" のデフォルトの訳語は「基礎」であるが、情報処理分野ではそのまま "basic" と訳出するのが適切である場合が多い。このような訳語指定は、他の分野でも多数みられ、それぞれの分野でのユーザから得られた有用な翻訳知識である。

5.3 動詞、形容詞等の他品詞に対する訳語指定

ユーザの訳語指定のインターフェースの簡便性のため、今回はユーザが訳語指定できる語は名詞のみとした。全ての訳語指定は名詞の訳語指定としてシステム側に解釈される。それにもかかわらず、ユーザが動詞、形容詞等の訳語指定を行なおうとする傾向が見られた。具体的には

adapted = 適応した
align = 整列させる
cooperable = 協力的

などの訳語指定が見られた。形容詞を無理に名詞として指定した場合には英語の「形容詞＋名詞」の単語列が「名詞＋名詞」の単語列として解釈されて正しい翻訳結果が得られる場合もあるが、動詞を無理に名詞として指定した場合には翻訳結果が悪化する場合が多い。

今後はユーザインターフェースの簡便性を考慮し、より多くの種類の品詞に関して訳語指定を行なえるようにする必要がある。ただし動詞に関してはその動詞固有の格パターンなども同時に指定しないとよい翻訳結果が得られない場合もあり、ユーザインターフェースに関する検討が必要である。

5.4 複数形の訳語指定

名詞の訳語指定は原形で行なうのが原則であるが、ユーザが複数形等の変化形で名詞を訳語指定してしまう場合が見られた。例えば

```
plug-in = プラグイン
plug-ins = プラグイン
gateway = ゲートウェイ
gateways = ゲートウェイ
```

と単数複数両方で指定されている場合などがあった。この訳語指定データの問題点は、

```
These plug-ins are ...
```

という文に対して、plug-ins 自身がシステムにとって単数形と解釈されてしまう可能性があるので、動詞 (are) と数の一致が得られず、解析に失敗してしまい翻訳結果が悪化することである。

訳語指定を行なうホームページ上に、訳語指定はかならず原形で行なうように記述を行ないユーザに注意を促すことでこの問題は解決できると考えている。

5.5 競合する訳語指定

複数のユーザが同じ見出しに対して異なった訳語指定を行うものもかなりの数が見られた。表3の最初の例は"encoding" という見出しに対して、2人のユーザが「符号化」という訳語を指定し、1人のユーザが「暗号化」という訳語を指定した場合である。

表 3: ユーザによって異なる訳語

見出し語	訳語指定
encoding	符号化 (2 人), 暗号化 (1 人)
ground	グランド (2 人), 土地 (1 人)
specification	仕様 (3 人), スペック (2 人)

これらの例はお互い競合する訳語指定情報である。訳語が違っていても、分野も異なる場合には問題がないが、分野が同じ場合には矛盾する知識となる。

5.6 ユーザの勘違いによるスペルミス

```
ploxy = プロキシー    (正しくは proxy)
infomation = infomation
                (正しくは information)
```

上の例に示すように、明らかにスペルミスと思われる単語の訳語指定もあった。原文がノンネイティブによって書かれているなど、原文中のスペルミスが原因で正しく翻訳できなかった場合に、ユーザが原文を修正せずにそのスペルミスの語をそのまま訳語指定したものと思われる。

5.7 URL、メールアドレス等の翻訳不要指定のための訳語指定

URL やメールアドレス等をそのまま訳出するように指定したものもいくつかあった。URL やメールアドレスが文中にそのままの形で現われた場合には翻訳しようと試みてしまう。この問題点を回避するためにユーザが訳語指定を行なったものと思われる。

6 結論

ユーザがインターネットホームページ上で簡便に翻訳依頼をできるシステムで、ユーザの訳語指定データをサーバ側に蓄積し、それによってユーザからの翻訳知識の収集を行った。ユーザが翻訳に使用する専門辞書を選択することで、自動的にユーザの翻訳知識の分野分類を行い、その結果翻訳システムにフィードバックできる生データを自動的に収集することができた。

また、インターネット分野、情報分野などでは最新用語等のデータが目立った。今後の翻訳システムの新辞書リリースに効果的にフィードバックできそうである。

また現在はユーザの訳語指定の対象品詞として名詞しかないが、他の品詞についてもユーザに訳語指定のニーズがかなりあることがわかった。ユーザインターフェースの簡便性を考慮しながら、今後の検討材料とする。

参考文献

[1] http://mtave.softpark.jplaza.com/MTave/

[2] FLM ネットワーク翻訳サービス
http://trns.cab.infoweb.or.jp/