ヒューマンインタフェース 78－5
（1998. 5. 15）

インターネット翻訳サービスにおけるユーザ辞書構築

熊野 明 中山圭介
(株)東芝 研究開発センター
〒210-8582 川崎市幸区小向東芝町1
{kmn, keisuke}@eel.rdc.toshiba.co.jp

アブストラクト

機械翻訳用ユーザ知識を翻訳ソフトの実ユーザから収集するために、インターネット上で翻訳サービスを公開した。ユーザは辞書登録によって自分のユーザ辞書を構築することができる。辞書登録は翻訳要求時に指定可能だが、翻訳サーバの用語抽出機能によって提示された情報を再翻訳時に利用することもできる。サーバは原文を翻訳すると同時に、原文中の未知語や辞書登録されていない複合語を抽出し、辞書登録の候補としてユーザに提示する。9 か月間の無料サービスで、約 6,700 語の辞書データが収集できた。このうち約 10%は用語抽出機能の抽出情報を利用したものであり、この機能が有効に利用されていることがわかった。ただし、抽出用語全体に対する辞書登録語の割合は 2%にとどまり、ユーザに提示するデータの品質として改良する余地がある。ユーザ辞書構築機能を備えた翻訳サービスにおいて、この用語抽出機能が効率的に機能していることがわかった。

Building User Dictionary in the Internet Translation Service

Akira Kumano, Keisuke Nakayama
R&D Center, Toshiba Corporation
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210-8582
{kmn, keisuke}@eel.rdc.toshiba.co.jp

Abstract

We have provided an Internet translation service in order to collect the dictionary knowledge for machine translation (MT) from real users. Each user is allowed to build his/her user dictionary on the server. Once a user submits a translation, they can use the data extracted during translation in order to refine the translation output. The MT server identifies the unknown words and the compound terms in the original text, which are not registered in the dictionary, and then shows them to the users for dictionary registration. During a nine-month experiment, about 6,700 terms were collected. Of these terms, 10% were registered based on the extracted data, which shows the extraction function performed well. However, the rate of registration data to all the extracted data is only 2%, which indicates that this should be a priority for future improvement. We consider though that the function of word extraction worked effectively in the Internet MT service with user dictionary registration.

1. はじめに

インターネットの普及に伴い、各種のサービスが公開されるようになっている。検索サービスはその代表的な例であるが、機械翻訳の機能を提供する翻訳サービスがいくつか現れている。

この背景には、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という状況がある。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、ユーザ辞書データとして収集・蓄積する仕組みを構築し、インターネット翻訳サービス上で公開した。ここでは、ユーザが知識をできるだけ登録しやすくするために、翻訳エンジンに辞書登録候補用語を抽出する機能を実装し、その結果を翻訳結果(訳文)とともにユーザに提示した。ユーザはこの用語抽出データを参考にして、辞書登録ができる。翻訳サービスにおけるサーバとユーザとのインタラクションで、この用語抽出機能の効果を確認したので、報告する。

2. 翻訳サービスとは

2.1 パッケージソフトとの比較

翻訳パッケージソフトは、国内でも 20 種類以上が販売されている。ユーザはパッケージを購入し、パソコンにインストールして利用するが、ソフトのバージョンアップは一般に有料であり、その都度自分で行なわなければならない。

また、システム辞書以外の辞書知識は、ユーザが個々に蓄積する必要がある。(一部のソフトには、ダウンロードにより辞書データの追加が可能なものがある[1]。)

これに対して、インターネット翻訳サービス[1]は、ブラウザ上のインタフェースを利用して原文を入力し、サーバで翻訳された結果を、電子メールなどで得るものである。パッケージを購入する手間や費用は不要だが、有料サービスの場合、翻訳量に応じた料金を支払うのが一般的である。

インターネット翻訳サービスでは、翻訳エンジンはサービス提供者側のサーバにあるため、サービス提供者が頻繁にバージョンアップすれば、多くのユーザが容易に最新の翻訳ソフトを利用することができる。これは、ユーザにとって最大のメリットである。

2.2 ユーザ辞書の重要性

言うまでもなく、機械翻訳の出力は常に完全ではない。これを補う最も基本的な知識は、辞書データである。様々な話題がインターネット上を流れ、その中で新しい用語や専門用語が生まれる。また、以前から存在していたがこれまで取り上げらることのなかった会社名、人名、地名でも、ある報道から多くの人に知られることもある。

このような用語に関する知識は頻繁に更新する必要がある。新しい用語を総合的に収集するのは非常に困難であるが、翻訳を必要としているユーザはまさにその語に直面している。したがって、そのような実ユーザが蓄積しようとしているデータを集めることができれば、サーバ側で自動的に新語のデータを収集することができる。われわれは、このデータの流れをできるだけスムーズにすることにより、効率的な収集が実現できると考え、インターネット翻訳サービスを構築した。

2.3 システムの特徴

本サービスは、インターネット上のソフトウエア販売・サービス提供サイト、ソフトパーク[4]を通して、英日翻訳サービス *MT Ave* (Machine Translation Avenue)[5]として 1997 年 7 月 8 日から公開した。図 1 に、*MT Ave* の構成を示す。

ユーザは英文の翻訳を要求し、その日本語訳文をサーバがユーザに返送する。後述するように、翻訳だけでなく、ユーザ辞書に知識を登録することができる。

[1] FLM のサービス[2](英日、日英両方向)、JICST のサービス[3](日英翻訳)などがあるが、いずれもユーザ辞書の構築を支援したものではない。

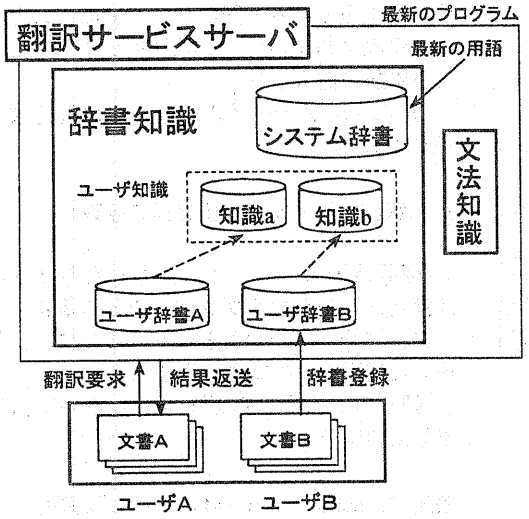

図1: 翻訳サービス MT Ave の構成

3. 翻訳サービスにおけるユーザ辞書構築

3.1 ユーザのメリットとサーバのメリット

ユーザは原文を入力する際に、名詞(単語、複合語)の訳語を指定することで辞書データの登録ができる。図2は ***MT Ave*** での入力例である。

machine = 計算機
translation software = 翻訳ソフト

図2：訳語指定の例

この例では、単語 machine に対して「計算機」、複合語 translation software に対して「翻訳ソフト」という訳語を指定している。

この機能は、従来の翻訳サービス[2]でも行なわれているが、***MT Ave*** では、これをユーザ毎のデータとしてサーバに蓄積することにより、ユーザ辞書が自動的に作成でき、以降の翻訳でも再度指定し直さなくても利用できる利点がある。もちろん、ユーザはいつでもその辞書データの内容を取り出すことができる。

また、サーバ側に蓄積されたユーザ辞書の内容は、翻訳ソフト開発者側では次バージョンの新しい辞書項目として、利用できる可能性がある。ユーザの辞書データをサービス提供者側が利用できるよう、本サービスの利用条件にそのための一項を含め、ユーザが ***MT Ave*** を利用する際には必ずこれを承諾していただいた。

なお、ユーザからできるだけ多くのデータを集めるため、今回は翻訳サービスを無料で開始した。これにより、ユーザにとっては、ユーザ辞書管理も含めた機械翻訳が無料で利用でき、サービス提供側では、実ユーザの記述した辞書データを利用できるという、双方にとってのメリットが生じる。

3.2 ユーザへの辞書情報提示

翻訳サービスの多くのユーザは、訳語指定の機能を理解しても、どの語に対して訳語を指定すれば効果があるのか、わかるとは限らない。

そこで、***MT Ave*** では、訳語指定することで訳文品質向上の効果が期待できる名詞(句)の候補を、英文翻訳時に抽出して、それらに対する現状での訳出結果とともにリスト型式にして、文書全体の翻訳結果とともにユーザに提示した。

具体的には、原文中から翻訳ソフトのシステム辞書に登録されていない単語・複合語(いずれも名詞)のうち一定頻度以上のものを抽出し、現時点での辞書知識による翻訳結果とともに示した。その抽出方法は、[6]で日英翻訳について行った方法を英日翻訳に応用したものである。

ここで抽出したものは、十分な知識で翻訳されていないと予想されるものである。したがって、ユーザが個々の用語に対する適切な訳語を辞書登録することにより、訳文の品質を改良させる可能性がある。

3.3 サーバとユーザのインタラクション

インターネット翻訳サービスにおけるユーザとサーバの間のデータの流れを、図3に示す。

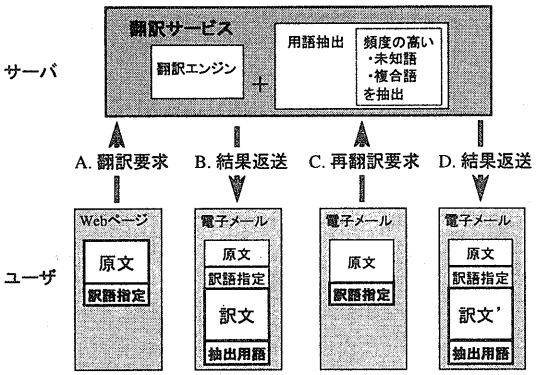

図３：ユーザとサーバのやり取り

A. 翻訳要求

文字通り、翻訳を要求する操作であり、WWW ページを通してユーザ側から行われる。

ユーザは翻訳する英文の他に、必要に応じて、(1)利用する専門辞書の分野、(2)訳語指定、などを入力する。これらのデータは、httpd サーバの cgi プログラムを通して翻訳サーバに送られる。翻訳サーバでは、入力された情報に基づいて、機械翻訳ソフトによって日本語に翻訳する。

図４は、本サービスで用いた翻訳要求ページである。

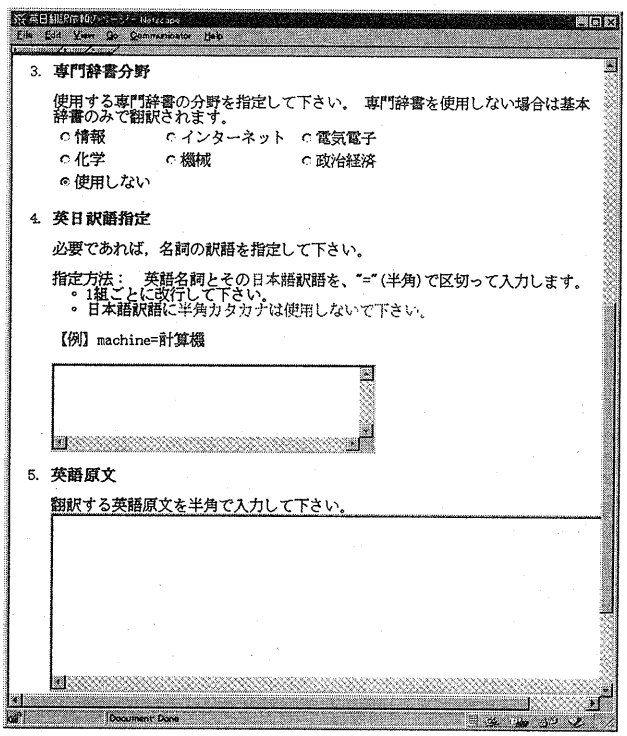

図４：翻訳要求画面

B. 翻訳結果・用語抽出結果の受取り

翻訳結果は、ユーザの入力した原文や訳語指定、その他のオプションなどの情報とともに、電子メールで送られる。ここには、前述した用語抽出の結果も含まれる。

図５は、訳文返送電子メール中の用語抽出結果の例である。

```
頻度 8: human translation = 人間翻訳
頻度 5: hostname = hostname
```

図５：返送メール中の抽出結果

最初の例は、human と translation という単語はシステムの辞書に入っているが、human translation という熟語としては入っていないことを示している。"=" の右の「人間翻訳」は、現状での翻訳結果である。human に対する「人間」という訳と translation に対する「翻訳」という訳を機械的に合成して出力されたものである。複合語の翻訳の場合、訳語の合成で訳出はできるが、この場合は「人手翻訳」という訳語を登録することで、訳文の品質を向上することが期待できる。「頻度 8:」は、この表現が原文中に現われた回数である。

2 つ目の例は、hostname という新しい語そのものがシステム辞書に入っていない場合である。まったく新しい語なので、"=" の右にある現状での翻訳結果も「hostname」となっている。

C. 再翻訳要求

返送された翻訳結果に対してより良い訳文を得ようとする場合、ユーザ

は、訳語指定に情報を追加して翻訳のやり直し、すなわち再翻訳を要求することができる。再翻訳は、翻訳結果メールに返信することで簡単に実行できる。

訳語指定を追加するには、用語抽出結果を参考にして、再翻訳要求の訳語指定部分に追加すればよい。図6は図5の抽出結果を利用した訳語指定の例である。

```
human translation = 人手翻訳
hostname = ホスト名
```

図6：抽出用語を利用した訳語指定の例

D. 再翻訳結果の受取り

再翻訳要求に対する翻訳結果は、最初の翻訳結果と同様に、電子メールで送られる。この結果に対して、さらに辞書データなどを変更して翻訳のやり直しを要求することもできる。

4. 辞書データの収集結果

無料翻訳サービスの公開を開始した97年7月8日から98年3月31日までの約9か月で収集したユーザ辞書データについてその特徴を分析した[1]。

4.1 分野別分類

収集された用語は、翻訳時に使った専門辞書の種類に応じて分野分類した。すなわち、翻訳時に専門辞書としてインターネット辞書を利用した場合の訳語指定は、インターネット分野の辞書データとした。分野分類されたデータの例を以下に示す。

- インターネット分野の例

```
Cool Site = クールサイト
data encryption = データ暗号化
Java platform = Java プラットフォーム
push technology = プッシュ技術
```

- 情報分野での訳語指定の例

```
bus reset state = バス・リセット状態
device driver = デバイス・ドライバ
method = メソッド
registry = レジストリ
```

このように、おおむね正しく分野が指定されていた。ユーザの入力した専門辞書の種類は、分野分類に利用できる見通しが立った。

4.2 見出し語と訳語の特徴

A. 大文字で始まる見出し語

原文全体に占める大文字単語の割合は少ないにもかかわらず、登録された単語の約3分の1が大文字で始まる単語に対する訳語指定であった。

```
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
```

これらの多くは地名、人名などの固有名詞であることから、新語として固有名詞の知識が必要とされていることがわかった。

B. 原語綴りのままの訳語

```
MIDI = MIDI
ping = ping
basic = basic
```

上の例のように英語の綴りをそのまま訳語に指定していした例が多数見られた。basic の標準の訳語は「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た。

この種のデータの多くは、用語抽出機能で抽出されたもの[2]をそのまま利用したものである。未知語を抽出した結果を修正することなく訳語指定する意味があると判断した例である。

[1] 詳細な分析結果の報告は [7]に譲る。

[2] 例えば、図5の hostname のような例

5. 用語抽出機能の評価

MT Ave の特徴である用語抽出機能が、ユーザの辞書データ構築においてどの程度貢献しているかを評価するために、2 種類の評価を行った。

A. 全辞書登録用語に対する抽出用語の割合

全ユーザが辞書登録したデータのうち、用語抽出データを利用したものの割合[1]を調べた。

ア。全ユーザの全辞書登録データ
6,662 語
イ。ア.のうち抽出用語に一致するもの
644 語 **(9.7%)**

辞書登録されたデータのうち、約 10%が、用語抽出機能の提示結果を利用したものと一致した。必ずしも用語抽出データを参考にしたとは言えないが、機械的処理で抽出したデータにもかかわらず、よく利用されたと言える。

B. 全抽出用語に対する辞書登録用語の割合

ユーザに提示した全抽出用語(辞書登録を 1 語も行っていないユーザのデータは除く)のうち、実際に辞書登録に利用されたものの割合[2]を調べた。

ア。辞書登録ユーザの全抽出用語
23,328 語
イ。ア.のうち辞書登録に利用されたもの
474 語[3] **(2.0%)**

辞書登録したユーザに対して提示した抽出用語のうち、実際に辞書登録に利用されたものは、わずか 2%であった。

[1] 用語抽出の精度を評価する観点からは、適合率に相当するものである。
[2] 用語抽出の精度を評価する観点からは、再現率に相当するものである。
[3] A.イ. には同一ユーザが同一抽出用語を異なる訳語で登録したものも別にカウントしているため、B.イ. の語数とは一致しない。

6. 結論

ユーザの辞書データ構築に、本サービスの用語抽出機能が利用されていることがわかった。特に辞書登録データの約 10%のデータは、用語抽出のデータを参考にしたものであり、用語抽出機能の効果が認められた。一般には辞書登録に際し、根拠とするデータを利用することは難しいが、本サービスの用語抽出機能によって、登録作業が軽減されていると考えられる。

これに対して、用語抽出データに対する実際の登録は約 2%にすぎなかった。現在、用語抽出には、簡単なアルゴリズムを利用しているだけなので、実際には翻訳に利用されないデータである記号類なども抽出・提示している。この処理方法を改良することで、抽出用語の利用率が高まる可能性がある。

今後は、さらにデータを分析し、インターネット翻訳サービスにおけるユーザ辞書構築の効果を明確にするとともに、本サービスの特徴である用語抽出機能の精度を向上させるよう改良していく予定である。

参考文献

[1] NEC, 翻訳アダプタ II 辞書の広場:
http://meshplus.meshnet.or.jp/adp2/dic/
[2] 富士通ラーニングメディア, ネットワーク翻訳サービス:
http://trns.cab.infoweb.or.jp/
[3] JICST, 翻訳ネットワーク:
http://www-jmt.jst.go.jp/
[4] ソフトパーク:
http://softpark.jplaza.com/
[5] 翻訳サービス *MT Ave*:
http://mtave.softpark.jplaza.com/MTave/
[6] 熊野 明, 平川秀樹: 対訳文書からの機械翻訳専門用語辞書作成, 情報処理学会論文誌 Vol.35, No.11, pp2283-2290 (1994).
[7] 中山圭介, 熊野 明: 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, 言語処理学会第 4 回年次大会, pp584-587 (1998).