機械翻訳システムの後編集ツールのススメ

武田 浩一
日本アイ・ビー・エム株式会社 東京基礎研究所
神奈川県大和市下鶴間 1623-14
takeda@trl.ibm.co.jp

1 まえがき

日本ではインターネット・ユーザのためのパーソナル機械翻訳(MT) ソフトが爆発的な普及をみせたことにより、従来のような専門性の高い文書の翻訳から、ブラウザでWEBページの概訳を生成することへと技術的な焦点が劇的にシフトしたといえる。価格や対応可能なブラウザ／HTMLのレベル、使い易さ、稼働環境、といった市場に受け入れられる製品そのものに要求される仕様を除けば、MTシステムが満たすべき主要な条件は、

1. 翻訳速度が速いこと
2. できる限り高い訳質であること
3. 単語／熟語登録などが容易であること

の3つといえる。翻訳速度はWEB翻訳という分野においては極めて重要であり、訳質とのトレードオフを考えると、「オンラインでの高速ナナメ読み」と「オフラインでのより高精度な翻訳」という使用形態に2分化しつつある。また、訳質は従来よりも「日本語としての自然さ」が重視され、誤訳を指摘するユーザの意見からは、誤った単語訳が予想以上に悪い印象を翻訳結果に与えるものと推定できる。学習機構を使用するユーザは全体からみれば少数ではあるが、学習により着実に正しい訳が生成できるという信頼感は高く評価される。より多くのユーザにこのような信頼感を提供できるためには、学習機構は使いやすくなければならない。現状では、翻訳速度は許容範囲、訳質はもう一歩、学習機構は可、という評価といえよう[7]。

MTシステムのユーザが100万人単位で存在すれば、個別に定義された単語／熟語／用例などの辞書が多くのユーザによって共有できることになり、このようなリソースの価値は従来よりも何倍にも高まるものと考えられる。従って、個別のMTシステムを上記の目標に向かって改良していくことはもちろん重要なことであるが、文書自体も含めた[8]共有可能なリソースの研究はそれに劣らず重要であるといえる。

ユーザ辞書共通フォーマット(UPF)[6]は、異なるMTシステム間でも辞書を共有できるように設計された辞書フォーマットである。ただし、このようなリソースを翻訳方式や翻訳知識に大きな違いのあるMTシステム間で共有するのは自明なことではない。例えば中間言語方式のMTシステムでUPFの単語辞書を利用するためには、意味分類などの限られた情報から概念要素への

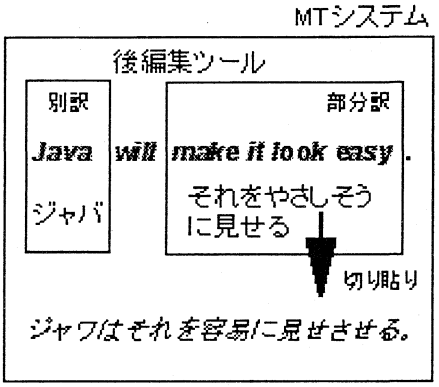

図 1: 後編集ツールの役割

写像を行なう必要があるし、既に非常に細かな意味マーカつきシステム辞書をもつMTシステムに、より大まかな意味分類しかもたない辞書を統合するのは容易ではない。

2 共有される翻訳知識

前述のように、異なるMTシステムでリソースを共有するという問題は、直接個別のMTシステムの翻訳処理に埋め込むという形では、短期間になかなか解決しないと予想される。従って、より現実的な方法として、入力文のタギングなどの前処理と、翻訳結果を補足する後処理において利用することを考える。前処理、後処理ともに、組み合わされるMTシステムとの独立性が高く、目標とする広い範囲のリソース共有が実現できる。前処理については既にMT以外の分野でも十分に研究されているため、本稿では、特に機械翻訳固有の後処理(後編集)における翻訳知識の共有について考察する。

後編集では、単語訳や用例のように、直接正しい訳を与える知識を適用しやすい。また、どのようなMTシステムでも原文と訳文は、ともに文字列という共通のフォーマットであるため、このレベルで共有リソースを用いた後編集ツールを提供できれば、翻訳方式とは独立して利用できる。直観的には、後編集ツールとは、図1のように、原文の理解を助けるような単語訳／部分訳を

表示し、その訳を切り貼りの形で利用することで、訳文の後編集を補助するツールであると考える。形式的には、カーソルの置かれた原文の単語に対して、その単語の別訳およびその単語を含む (様々な句／節の) 部分訳の集合を、それぞれの訳の重みまたは尤度による順序付けとともに計算するプログラムであると定義する。

このような後編集ツールには、様々な翻訳知識が応用できる。特に、最近のコーパスを利用した統計処理により、自然言語の様々な曖昧性解消が高い精度で行なえるようになってきた。特に最近の研究成果には注目すべきものが多く、語義推定 [5]、品詞推定 [2]、共起表現の抽出 [3] と単語訳の推定 [1] など、翻訳に直接適用できる手法が明らかになりつつある。このようなコーパス／統計手法から得られるリソースと、ユーザが学習機構を通して定義できる知識をシームレスに統合し、できる限り高い精度で単語訳や部分訳を計算できれば理想的である。このような翻訳知識として、

- 共起辞書:
  $\{r \mid r = sw_1, \ldots, sw_n \rightarrow tw_j (n \geq 1, 1 \leq j \geq n)\}$
- 対訳パターン:[4]
  $\{p \mid p = sw_1, \ldots, sw_n \rightarrow tw_1, \ldots, tw_m (n \geq 2, m \geq 1)\}$

という 2 種類の知識を提案する。共起辞書は、原言語文で共起する n 個の単語 $sw_1, \ldots, sw_n$ が存在するときに、j 番目の単語 sw_j の訳が tw_j であるという知識の集合である。$n = 1$ の場合を許すので普通の単語辞書を完全に含んでおり、共起する単語の並びや、文頭などの特定の位置にある単語が現れるといったに制約を記述するような拡張も可能である (ただし右辺に複数個の単語に対する訳語が指定できるようにしても表現能力は変わらない)。対訳パターンは、原言語文でこの順に現れる n 個の単語 (または文節)$sw_1, \ldots, sw_n$ からなる表現に対して、その対訳が $tw_1, \ldots, tw_m$ であるという知識である。各知識には重みが定義されていてもよい。このような知識は、単純な形式をしており、コーパスから統計的処理によって獲得したり、ユーザが直接定義したりすることが可能である。さらに、この 2 種類の翻訳知識は、以下のような豊富な特徴を有する。いま簡単のために規則の重みや優先度は考慮せず、2 つの共起辞書 D_1, D_2 と、2 つの対訳パターンの集合 P_1, P_2 があるものとする。

- 同じ単語に対して別の訳を定義する 2 つの共起辞書規則に 3 種類の競合が定義できる。一方の左辺に現れる単語の集合が、他方の集合を真に含む時は「例外」、両者の集合が一致する時は「別訳」、両者が比較不能の場合は「訳し分け」と呼ぶことにする。これ以外に規則の競合は発生しない。同様の競合は対訳パターンについても定義できる。さらに、特定のテキストの上で競合する規則と、潜在的に競合する可能性のある規則が考えられる。
- 左辺の単語集合が包含関係にあり、右辺の単語訳が同じ 2 つの共起辞書規則に冗長性が定義できる。任意の共起辞書や対訳パターン集合に対し、非冗長で最小のものが 1 つ存在する。
- D_1, D_2 と、それらが定義する訳の集合は、集合和について閉じている。また、その和が非冗長であるなら、集合積についても閉じている。P_1, P_2 の場合は、常に集合演算に関して閉じている。

このような競合と冗長性の概念は、複数の翻訳知識を共有し最適な翻訳知識を合成する場合に極めて重要な役割を果たす。

3 今後の研究

本論文では、後編集ツールのアイデアと、そこで利用可能な翻訳知識の基本的な性質について考察した。このような翻訳知識を用いて、正しい第一順位の訳を求めるには、規則の優先度／重み付けが不可欠となる。これについては、コーパスに現れる頻度から重みを計算する様々な手法や、コーパスを構成する文書集合がわかる場合には、階層クラスタリングによって分類し、1 つのコーパスを、あたかも個別の専門分野が存在するかのように文書部分集合を求めて、そこから翻訳知識の部分集合を計算する手法などが有望である。

参考文献

[1] I. D. Melamed. "A Word-to-Word Model of Translation Equivalence". *Proc. of ACL/EACL'97*, pp.490–497, Jul. 1997.

[2] H. Schütze and Singer Y. "Part-of-Speech Tagging using a Variable Memory Markov Model". *Proc. of ACL'94*, pp.181–187, Jun. 1994.

[3] S. Shimohata, T. Sugio, and J. Nagata. "Retrieving Collocations by Co-occurrence and Word Order Constraints". *Proc. of ACL/EACL'97*, pp.476–481, Jul. 1997.

[4] K. Takeda. "Pattern-Based Machine Translation". *Proc. of COLING'96*, pp.1155–1158, Aug. 1996.

[5] D. Yarowsky. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora". *Proc. of COLING'92*, pp.454–460, July 1992.

[6] 伊藤他. "機械翻訳ユーザ辞書の共通フォーマットの設定". 言語処理学会 第 *3* 回年次大会, 1997 年 3 月.

[7] 塩田. "パーソナル英日翻訳ソフトウェア". 月刊スーパーアスキー, 8(9):121–152, 1997 年 9 月.

[8] 橋田他. "大域文書修飾: 標準タグによる言語データの大規模な構造化と再利用". 言語処理学会 第 *3* 回年次大会, 1997 年 3 月.