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ることにより、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の技術的な焦点が割り合いシフトしたといえる。価格や対応可能なブラウザ/UIHTMLのレベル、使いやすさ、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マ読み」と「オンラインでのより高精度な訳し」という使用形態に2分化しつつある。また、訳質は従来も「日本語としての自然さ」が重視され、訳訳の指摘するユーザの意見には、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。学習機構を用いるユーザは全体からみれば少数であるが、学習により着実に正しい訳が生成できるという信頼性は高く評価される。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、学習機械は使いやすくなければならない。現在では、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。

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ユーザによって共有できることになり、このような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が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形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ではなくない。例えば対中間言語方式のMTシステムでUPFの単語辞書を利用するために、意味分類などの限られた情報から概念要素の

MTシステム

後編集ツール

別訳
Java
シェバ
will
make it look
easy.
部分訳
それぎやししば
に見せる
切り貼り
シェバはそれを容易に見せさせる。

図1: 後編集ツールの役割

写像を行う必要があるが、既に非常に細かな意味マー
カつき辞書辞書をもつMTシステムに、より大ま
かな意味分類しかもない辞書を組合せるのは容易で
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なく、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、入力文のタグインなどの前処理と、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、組み合わせられた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。

後編集は、単語辞書用例を直接ある訳を与える知識を適用しよう。また、どのようなMTシステムで原語と訳文は、とともに文脈と文脈が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作成できれば、翻訳方法を共に独立
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辞書/部分辞を

—592—
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、カーネ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文節/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。


- 共起辞書：
  \( \{ t | t = s_{1}w_{1}, \ldots, s_{n}w_{n} \rightarrow t_{j}(n \geq 1, 1 \leq j \leq n) \} \)

- 対訳パターン：
  \( \{ p | p = s_{1}w_{1}, \ldots, s_{n}w_{n} \rightarrow t_{1}, \ldots, t_{m}(n \geq 2, m \geq 1) \} \)

という2種類の知識を提案する。共起辞書は、原語文で共起する2個の単語 \( s_{1}w_{1}, \ldots, s_{n}w_{n} \)が存在するときに、 \( j \)番目の単語 \( s_{j}w_{j} \)の訳が \( t_{j} \)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、文脈などの特定の位置にある単語が現れるといった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、原語文文でこの順に現れる2個の単語（または文節）\( s_{1}w_{1}, \ldots, s_{n}w_{n} \)からなる表現に対して、その対訳が \( t_{1}, \ldots, t_{m} \)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る。このような知識は、単純な形式をとることがあるが、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種類の翻訳知識を“パターン”として、より特定の特徴を有する。"同様の文脈"のように、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訳を定義する2つの翻訳辞書規則に3種類の競合が可能である。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を考える。

- 左辺の単語集合が含む領域があり、右辺の単語群が同じ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。 \( P_{1}, P_{2} \)の場合同値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1順位の訳を求めることは、規則の優先度と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分野を集めて、そこから翻訳知識の部分集合を計算する手法がある。

参考文献


