

機械翻訳と訳例検索を統合した翻訳支援システム

潮田明、富士秀、大倉清司、山下達雄

(株) 富士通研究所

ushioda@jp.fujitsu.com

1. はじめに

機械翻訳技術と訳例検索技術を統合した翻訳支援システムを開発し、性能評価を行ったので報告する。高速で汎用的であるという機械翻訳の長所と、分野固有の表現や慣用句を人間が柔軟に使い回しできるという訳例検索（翻訳メモリ）の長所を組み合わせることにより、人手翻訳と同等品質の翻訳文を従来の人手翻訳より格段に効率的に作成できることが確認できた。本システムを用いた評価実験を行ったところ、WEB上のオンラインヘルプマニュアルの英日翻訳で人手翻訳の約3倍、特許文書の日英翻訳で約4倍の効率化が図れた。なお、実験に使用した訳例は、テスト用翻訳文と分野は同じだが出所は全く別のものであり、従来の翻訳メモリより汎用性が高いことも明らかになった。以下の節ではまず翻訳支援システムの構成と翻訳作業の流れについて述べ、続いて各構成要素を説明し、最後に本システムの評価結果を示す。

2. 翻訳支援システム Cliché

機械翻訳の研究の歴史は古く、コンピュータが誕生した当初から機械翻訳には大きな期待が寄せられていた。わが国においても1970年代末より産業翻訳の効率化を目指して商用機械翻訳システムの研究開発が行われてきた。しかし20年以上たった今日でも一般ユーザが十分満足行く性能レベルには至っておらず、翻訳業界における機械翻訳の浸透度もまだ低い。プロの翻訳者が機械翻訳をまださほど積極的に使おうとしない最大の理由は翻訳品質が十分でないことであると思われるが、それと同じぐらい重要な要因として、機械の後処理を人間が行うという機械優位の作業方法に対する抵抗感がある[1]。一方で、技術的には機械翻訳よりも遥かにシンプルな翻訳支援システムは着実に翻訳業界の一部に足場を築きつつある。しかし、現在主流の翻訳支援システムである翻訳メモリも、用途がマニュアルなど過去の翻訳文の繰り返しが多い分野に限られ、産業翻訳全般に役立つ技術とはなっていない。

そこで、我々は今回機械翻訳技術と訳例検索技術を統合した新しいタイプの翻訳支援システムを開発し、性能評価を行った。基本はあくまで人間による翻訳であり、ユーザがすべてのコントロールを握るため上述のような抵抗感は少ない。一方でユーザの翻訳作業における各工程を極力省力化するように既存の技術をツール化して提供している。図1に今回開発した翻訳支援システム Cliché の構成図を示す。クライアントシステムは翻訳対象文から原文を1文ずつ切り出して翻訳エディタに提示する。翻訳者はエディタ上で訳文を作成するが、その際 GUI を通して、機械翻訳サーバおよび訳例検索サーバにアクセスし、原文やその一部に対する自動翻訳結果や訳例検索結果から有用な表現を選択し作成中の訳文に簡単な操作で挿入することができる[5]。

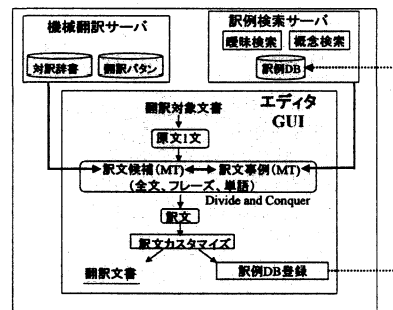


図1: 翻訳支援システム Cliché の構成図

訳文作成の手順は翻訳者の自由であるが、多数の翻訳者による試験運用から分かった効率的な手順は以下の通りである。

まず基本的な方針は“divide-and-conquer”である(図2)。これは Example-Based MT [2,3]における機械処理の理想的な工程に似ている。まず文の大まかな構造をつかみ訳文の骨格を決める。そして骨格の主要要素をそれぞれ翻訳し、最後に各要素の翻訳を骨格にはめ込んで訳文を完成させる。要素自体が長い、あるいは複雑な場合はその要素を更に骨格と要素に分割して翻訳を進める。このように各要素を分割していくと、ある時点で同一の表現が訳例中に見つか

るか、あるいは機械翻訳（部分翻訳）で正しく訳されるようになる。作成された訳文は最後に訳文スタイルのユーザカスタマイゼーションが施されて翻訳文書に挿入される。このような翻訳の進め方は、特許などのように一文が長い文書に特に有効である。分割した後は他の分割要素とはある程度独立して訳せることが必要になるため、divide-and-conquer approach がすべての文書に有用であるわけではないが、機械とは異なり人間が訳す場合は各要素間の繋がりを頭に留めながら翻訳が行えるため、独立の条件は緩和される。

原文:

〔本発明により提供された統合型光送信装置〕は、〔光ビームを受光しそれに対し変調信号に応じ変調を与える光変調器〕によって特徴付けられる

骨格文: ①本装置は、②光変調器 によって特徴付けられる。

→ ①This device is characterized by ②an optical modulator. (機械翻訳)

①本装置	→	本発明により提供された統合型光送信装置
① An integrated optical transmitter provided in accordance with the present invention [訳例検索]+[機械翻訳]+[手編集]		
②光変調器	→	光ビームを受光しそれに対し変調信号に応じ変調を与える光変調器
② an optical modulator which receives the optical beam and provides modulation thereto in response to modulation signals [訳例検索]+[機械翻訳]+[手編集]		

最終訳文

[An integrated optical transmitter provided in accordance with the present invention] is characterized by [an optical modulator which receives the optical beam and provides modulation thereto in response to modulation signals.]

(a) 骨格文-主要素の分割と翻訳結果の結合

- ② 光ビームを受光しそれに対し変調信号に応じ変調を与える光変調器
- ②-1 光ビームを受光し receives the optical beam [訳例検索]
 - ②-2 変調信号 modulation signals [機械翻訳]
 - ②-3 変調を与える provides modulation [訳例検索]
 - ②-4 光変調器 an optical modulator [機械翻訳]

⇒ ② an optical modulator which receives the optical beam and provides modulation thereto in response to modulation signals [手編集]

(b) [訳例検索]+[機械翻訳]+[手編集]の例

図 2: Cliché における翻訳作業の流れ

3. 訳例検索

Cliché の解析ツールの核の一つが訳例検索である。翻訳者はエディタ上の任意の連続文字列をマウスで選択し、クリック操作で訳例検索を行うことができる。訳例データの検索においては「絞り込み」と「マッチング」の 2 段階の処理が行われる[6]。絞り込み検索では、suffix array[8]により、指定さ

れた絞り込み数分の結果を得る。この後、絞り込まれた検索結果の一つ一つを検索キーワードとつきあわせ、ダイナミックプログラミング法により最適な照合結果を取りだすアルゴリズムにより単語同士の対応を得る。検索結果は、複数の対訳データが類似度順に表示される。類似度計算方式は、「一致率」や「スコア」などが選択でき、その場で即座にランキング表示に反映させることができる。

一致率は、検索キーワードと訳例原文の文全体の一致を重視するもので、百分率などで表される。完全に一致する場合は 100% となり、これがこの類似度の上限となる。従来の翻訳メモリシステムの多くはこのタイプの類似度評価を行っている。一方、スコアは、検索キーワードと訳例原文との部分一致を重視するものである。単語が連続で一致している場合に得点を付加するなどの方法で、連続部分での一致が多い訳例に高い類似度を与える。契約書の訳例データベースを用いた英日翻訳実験においてスコアと一致率のどちらの評価値が実際の翻訳作業に有効か評価を行ったところ、全般にスコアの方が有効であるという結果が得られている[6]。

Cliché の訳例検索機能のもう一つの特徴は、訳例データの構造にある。訳例は従来の翻訳メモリシステムのように単に日英 2 ヶ国語ストリングのペアとしてデータベースに格納されているのではなく、英日・日英翻訳ソフト ATLAS[4]の解析エンジンによって得られた日英文の解析情報がストリングに付加された形で XML 形式で格納されている。Cliché の訳例検索モジュールにおいては形態素解析情報をはじめ、構文構造や意味構造を用いた訳例検索も可能な機構になっている。すなわち訳例検索に機械翻訳技術が埋め込まれていると言える。

図 3 に訳例データベース作成の流れを示す。訳例の作成は、訳例データベースを作成する際のみでなく、翻訳中に出来上がった訳文を原文とペアにして訳例データベースに登録する際にも行われる。訳例の作成においては、各言語を構成する単語同士の対応が、意味情報を用いて、あらかじめ自動的に付与される。

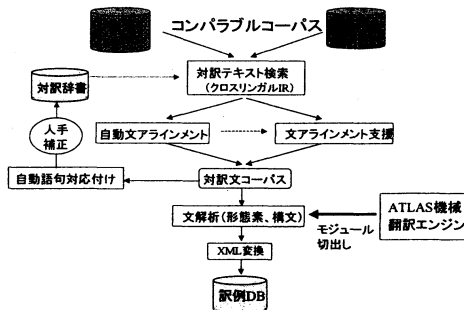


図 3: 訳例データベースの作成

それぞれの言語の文を構成する単語列を、両言語共通の意味タグ列に変換し、意味タグ同士のマッチングをとることで単語対応が図られる。この機能により検索された日英間で対応のついた語句がハイライト表示されるので、翻訳したい語句をすぐに見つけることができる。これらの訳例データは、言語情報・対応情報とともに、XML に準拠したフォーマットの訳例データベースファイルへ格納される。訳例データベースファイルは、検索で利用されるのはもちろん、リアルタイムでの訳例データの追加・削除が可能である。

Cliché の訳例検索機能の 3 つ目の特徴は、ネットワークを介した訳例データベースの共有が可能なことである。インターネットあるいはイントラネットを通じて、複数の翻訳者がリアルタイムに権限の与えられたデータベースに対して対訳文の参照/登録/削除が行える。プロジェクトマネージャーがそれらを統合して、訳例の統一をはかることもできる。この共有機能は遠隔地でのグループ翻訳作業において特に有用である。

4. 機械翻訳

Cliché の機械翻訳モジュールには英日・日英翻訳ソフト ATLAS[4]のエンジンをを用いている。比較的単純な文については、機械翻訳結果を少し手直して訳文を完成できるが、長い文や構文の複雑な文においては構文解析の失敗するケースが多く、人手による修正が、一から翻訳するより煩わしくなることも珍しくない。このような文に対してはこれまでの翻訳現場では機械翻訳は全く貢献できないことが多い[1]。しかし Cliché においては、翻訳文全体は利用できなくても、自動辞引き機能として有効に使うことができる。ユーザ辞書にマッチした単語やフレーズがあった場合は訳語が強調表示されるので、用語集などをユーザ辞書にあらかじめ入れておくことにより、信頼できる訳語が一括して表示される。また訳語に誤りがある場合、訳語候補を表示、選択することができ、その結果を学習させることができる。

Cliché においては訳例検索に機械翻訳技術が使われているだけでなく、機械翻訳にも訳例データベースからのフィードバックがかけられる。図 3 に示すように、訳例データベースを作成する際に文単位でアラインメントされた対訳文コーパスが作成されるが、対訳文コーパスから同時に自動対訳語句抽出システムと人手修正により対訳辞書が作成され、機械翻訳の用語集として使用される。従って、同一分野での訳例が蓄積されるに連れて対訳辞書も整備され、機械翻訳の精度向上へとフィードバックされていく。また、機械翻訳の精度が向上すれば、訳例の解析精度も向上するため、検索の精度も向上する。このようなループは、特にグループで翻訳を行う際に効果が期待できる。

5. 評価実験

WEB上のオンラインヘルプマニュアルの英日翻訳および、特許明細書の日英翻訳において、Cliché を用いた場合(Cliché 翻訳)と、パソコン上での従来の翻訳環境で翻訳を行った場合(手翻訳)とで翻訳の精度と速度にどのような差が生じるか評価を行った。ここでいう手翻訳には、オンライン辞書やインターネット検索の使用が含まれる。なお今回の実験では、翻訳者による効率のばらつき度合いを確認する意味から、訳例データベースの共有は行わなかった。

オンラインヘルプマニュアルの英日翻訳実験では、翻訳会社の翻訳者 2 名 (A, B) が被験者となり、実務での翻訳依頼に対し、手翻訳と Cliché 翻訳を行った。2 名の翻訳者は同一の文書に対して手翻訳を行い、また別の同一文書に対して Cliché 翻訳を行った。実験に使用した訳例データベースは、オンラインヘルプマニュアルの対訳文書から作成したものだが、評価用に用いたオンラインヘルプマニュアルとは別の会社のものであり、従来の翻訳メモリシステムの使用環境(マニュアル改版や Web のアップデート)での設定とは大きく異なる。

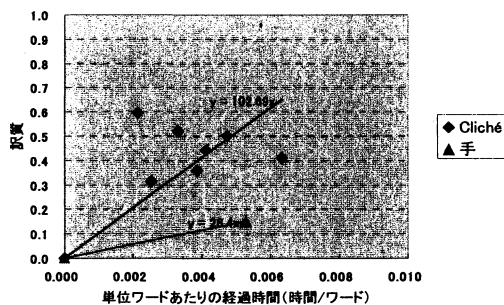


図 4: 翻訳者 A の翻訳効率化

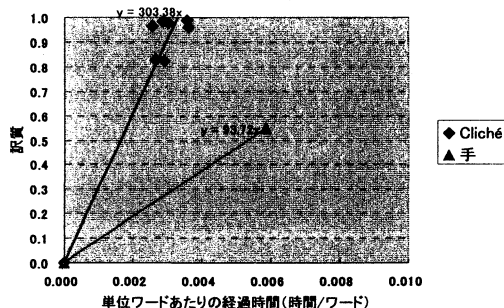


図 5: 翻訳者 B の翻訳効率化

図 4~5 に評価結果を示す。図は手翻訳および Cliché 翻訳について、単位ワード当たりにかかった翻訳時間と翻訳結果の訳質の関係をそれぞれの翻訳

家についてプロットしたものである。当初は単に翻訳スピードのみの評価を行って比較していたが、翻訳者によっては分からない箇所はあまり時間をかけずに進んでしまいスピードは速いが訳質は不十分といったケースが多々あるため、スピードだけでは適正な評価が難しかった。今回の評価法を導入することにより、より厳正な評価が可能になった[7]。

訳質の評価は、翻訳会社内での評価基準に従い訳質をAからCの8段階に分けて8点満点評価を行い、それを100%評価に換算することで行った。図中の直線は、原点を通る近似直線である。図において一定訳質の直線（水平直線）と近似直線との交点でのx座標は、その訳質に到るまでかかる1ワード当たりの翻訳時間を示す。従って、2つの近似直線の傾きの比を求めれば、一定訳質に到るまで翻訳を続けるという条件のもとでの、Cliché翻訳と手翻訳の翻訳速度の比（翻訳効率）が求まる。二人の翻訳者の翻訳効率自体は大きく異なるが、図4、図5ともCliché翻訳の効率は手翻訳の約3倍であることを示している。ただし、図4～5の評価実験では、実務と評価を兼ねていたため、効率の悪い手翻訳のデータは1ポイントしか得られなかった。効率化の正確な数値を得るためには手翻訳のデータポイントを増やす必要がある。

図6は、特許明細書の日英翻訳における評価結果である。この実験ではホームページ翻訳のときは別の2人の翻訳者(C, D)を被験者としたが、2人とも特許翻訳の経験はない。評価基準は図4～5のものと同様であり、誤訳を含まない文の数の比率を基準に求めたが詳細は[7]に譲る。図6では、2人の翻訳者の効率はCliché翻訳、手翻訳とも似通っており、2人ともCliché翻訳の効率は手翻訳の約4倍という結果が得られた。

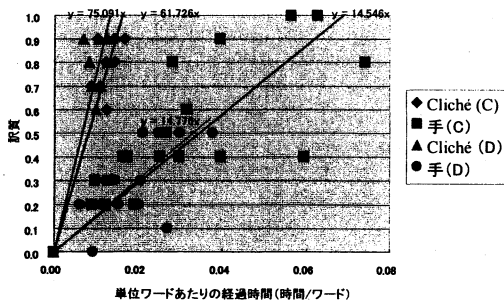


図6: 翻訳者C, Dの翻訳効率化

6. まとめ

機械翻訳(MT)と訳例検索(TM)を統合した翻訳支援システムを開発し、性能評価を行った。本システムにおけるMTとTMの統合はまだ発展途中の段階というべきであるが、以下の3つの特徴を備えている。まず、専用エディタ上で、原文の任意の文字列に対

して、MTとTMがワンクリックで呼ばれ、機械翻訳結果と訳例検索結果がつなぎ合わされて翻訳文が完成していくという意味でMTとTMが連携している。次に、訳例データにはMTによる解析情報が付加されており、この情報を用いて検索や表示が行われる。そして、蓄積された訳例から対訳情報が抽出され、機械翻訳にフィードバックがかけられる。divide-and-conquer approachによる翻訳は現時点ではエディタ上での人手による編集が主体となるが、今後編集のための自動システムの開発を進める。すなわち、原文を統語的あるいは意味的にまとまりのある要素表現に分割し、各要素表現の翻訳を行った後に原文の分割過程を逆に辿りながら各要素表現の翻訳を繋ぎ合わせて翻訳文を完成させる際に、分割や翻訳の履歴や繋ぎ合わせの手順を人間が意識することなく全体の翻訳を完成させることができるシステムを目指していく。

現時点でのMTとTMの統合の効果を評価するために、本システムを用いた翻訳効率化の測定を行ったところ、WEB上のオンラインヘルプマニュアルの英日翻訳で人手翻訳の約3倍、特許文書の日英翻訳で約4倍という結果が得られた。また、実験に使用した訳例が、テスト用訳文と分野は同じだが出所は全く別のものでも有用であり、従来の翻訳メモリより汎用性が高いことも明らかになった。

謝辞 Cliché開発の段階からClichéの評価に全面的にご協力いただいた(株)十印に感謝いたします。

参考文献

- [1] Brinkmann, Karl-Heinz.(1980) "Terminology Data Banks as a Basis for High-Quality Translation." Proceedings of the 8th International Conference on Computational Linguistics.
- [2] Makoto Nagao (1984) "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle." In Alick Elithorn & Ranan Banerji, eds., Artificial and Human Intelligence, Elsevier Science Publishers, chap. 11, pp. 173-180.
- [3] Satoshi Sato and Makoto Nagao (1990) "Toward Memory-based Translation." Proceedings of the 13th International Conference on Computational Linguistics, pp. 247-252.
- [4] 富士通. 英日・日英翻訳ソフト ATLAS. <http://software.fujitsu.com/jp/atlas/>.
- [5] 大倉清司, 山下達雄, 富士秀, 潮田明. 機械翻訳と訳例検索を統合した翻訳支援システムのインターフェース. 言語処理学会第9回年次大会予稿集, 2003.
- [6] 山下達雄, 富士秀, 大倉清司, 潮田明. 翻訳支援に有効な訳例検索の類似度計算方式と検索結果提示方式. 言語処理学会第9回年次大会予稿集, 2003.
- [7] 富士秀, 潮田明, 大倉清司, 山下達雄. 翻訳支援システム導入による効率化の評価. 言語処理学会第9回年次大会予稿集, 2003.
- [8] Udi Manber and Gene Myers. (1990) "Suffix arrays: A new method for on-line string searches." 1st ACM-SIAM Symposium on Discrete Algorithms, pp.319-327.