機械翻訳用テストデータ

5R－9

田中康仁
兵庫大学
E－mail:yasuhito@humans－kc.hyogo－dai.ac.jp

〔０〕はじめに

機械翻訳システムが日本に於いて本格的に研究開発されはじめて約２０年弱の年月が経過した。しかし、機械翻訳には色々な問題点がある。ここでは機械翻訳の現状を概観し品質向上にあたっての問題点、方法を検討する。

〔１〕機械翻訳システムの現状

１）機械翻訳システムの開発会社

日英、英日の機械翻訳システムの開発会社は十数社になる。このほか、自社業務用に開発したところを加えるとまだ増える。また、機械翻訳システムの中心部分をＯＥＭし、製品化している会社、専門用語辞書をもっぱら作成している企業、販売している企業もある。

２）研究開発を続けている企業

これらの企業の中でも、研究部門は既になくなってしまったところ、開発も少人数でしか行っていないところも出現している。このあたりも詳しく調べておかなければ使用者は不安の残るところである。

３）専門用語辞書

企業の供給している専門用語が少しずつ、古くさくなり最新の技術用語が組み込まれていないという問題が発生し始めている。これは約２０年弱前に研究開発した頃、専門用語を作成した、またＯＥＭで購入し組み込んだままで、その後更新を行っていないためである。普通の一般用語辞書についても同じような問題点がある。

４）価格

ワークステーション上の翻訳システムは数１０万円、専門用語辞書もそろえると百万円程度かかるというものが、このごろではインターネット上の英文のおおまかな翻訳システムは数万円になっている。

利用者の急増とパーソナル・コンピュータの性能の急速な向上と普及により、機械翻訳システムは数万円程度にまで安くなってきた。これは利用者にとって大変喜ばしいことである。しかし、一部の企業は専門用語辞書の価格はあまり安くしなくて、全システムの価格は安くならないようにするとか、専門用語辞書の導入を１分野とか数分野に制限する等で価格の下落を防ごうとしている。しかし、全般的には安く導入できる方向に向かっている。

５）翻訳速度

パーソナルコンピュータの性能の向上にともない処理文数が２万文程度でもペンテイアムⅡ３００ＭＨｚの機械で１時間強で処理が終るまでになっている。翻訳速度は大きな問題ではない。

６）品質

機械翻訳システムで最大の課題は品質である。まだまだ十分なものには到ってはいない。これは開発者の問題であると同時に利用者からの問題提供、翻訳の質に対する系統的なクレームの申立てによる改良にあると思われる。利用者が商品に対して申立てを行うのは当然のことであるがこれが充分行われていないのではないだろうか？雑誌等の評論記事はソフトウェア会社や開発企業の宣伝に振りまわされていて本当の姿が表現されていない。

日本電子工業会を中心にした企業の研究会で、機械翻訳の機能テストを行っているが、プログラムテストの観点からは有効であるが、機械翻訳システムが持たなければならない知識の量については充分な検査とはなっていない。

今では機械翻訳システムにどの程度有用な知識データが入っており、それが活用されているかということを調べる段階にきている。

数文程度の比較ではなく数万件程度のテストデータを常に準備する。しかも、半年毎とか、１年毎にそれらを更新して、新しいテスト文で検査して、改善を計らねばならない。日英、英日の検査を行うことを考えると、パラレル・コーパスの作成が必要である。

消費者センターと協力し、品質の向上に努めることが重要である。

機械翻訳システムには限界が明確に述べられていない。また、どれだけの文でどのようにテストしたかについても述べられていない。これは利用者の誤解と失望をまねく。

さらに必要なことは、今後どのような方針で改定をするかということも書かれていない。閉じた完全なシステムであるならばよいが、自然言語のようにオープンな世界で利用するものには記述が必要である。

７）市場占有率

機械翻訳システムがパーソナルコンピュータ利用者にどの程度普及しているか？　どこの会社の製品がどの程度の市場を占めているか、機械翻訳システムをどのような目的のために利用しているか等、利用動向を調べる段階に来ている。

８）機械翻訳システムの版（ヴァージョン）

機械翻訳システムは第○版というように表現されている。版が変わるごとに大きく改良されて使いよいものに変わっている。現在のものは数多くの改変は進められいるが、第３版とか第４版程度である。

ワードプロセッサーのソフトウエアで有名な某社のものは第８版にもなっている。辞書は１２回の改良がなされている。これは日本語一つの処理でこの程度まで改良されやっと使いよいものになっている。一般的にソフトウエアの改版は１～２年に一度行われる。

Test Data for Machine Translation System.
Yasuhito Tanaka
Hyogo University

このことから考えるとあと8回程度改版されるとすると約10年～15年後にはかなり良い機械翻訳システムが市場に出回ると推測する。

〔2〕どのようにしてテストデータを作るか？

テストデータを作るにあたっては次の三つの方法が考えられる。

1）大量のコーパスを分析する。

2）既に出版されているCD－ROMの利用

3）多勢の人によるテスト文の作成

等が考えられる。次にこの個々の方法を詳細に述べる。

1)の方法として英語の新聞その他のデータ・ベースが機械可読媒体として売られている。これらは研究等については特別の許諾書に著名をすれば利用可能である。このようなものは多くの場合CD－ROMで売られている。量としても適当である。

2）の方法として既に出版されているCD－ROMを利用する方法としては

英和辞典、和英辞典として売られているCD－ROMをDDWIN32等の検索ソフトを利用し、例文を抽出し利用することも考えられる。英文のビジネス文作成のためのCD－ROMが売られている。これらを利用するのも一つの方法である。

しかし、著作権の問題も考えなければならない。

3）の方法としてあげた多勢の人々によるテスト文の作成は我々のように大学で教えているものにとってはデータ収集のよい方法である。

100人の学生のタッチ・タイプの練習として英文、日本語文の対を入力させる。約300文を入力する。前期、後期では2回の機会があり、3万文×2＝6万文（英⇔日の文）が機械可読媒体で集まる。しかし、学生の雑多な文をうまく整理する方法を考えなければならない。あまり複雑な入力方法では誤った文ばかりになる。

大量に集められた文を最初から翻訳システムに利用するのも一つの方法であるが次のような処理が必要である。

1）文の切り出し、文の認定、文の検査。

2）文の構成単語数（英文の場合）で分類する。

3）同一の単語数の文の中で、単語数が1～10単語の文の中には同一の文が複数件ある。これは例文翻訳としても採用できる。

このようにして文を単語数で整理すると文の翻訳の困難さの程度を容易に判断することができる。翻訳処理の1つの単位として取り扱いやすい。

次に日本電子化辞書の英文コーパスを単語数で分析した。

単語数	文数
1	0
2	19
3	460
4	1,889
5	5,030
6	6,798
7	7,791
合計	21,987

〔3〕テスト手順

次のような手順で機械翻訳システムをテストする。

1）7単語までの文で約2万件のデータが得られる。このデータを機械翻訳システムにかけて翻訳する。

2）翻訳結果に対して良否の評価を行う。1～5点の点数を付ける。

3）点数別、単語数別に分類する。同一点数で翻訳に悪い影響を与えている原因コードを付ける。複数の原因コードがある場合は1つだけ原因コードを付ける。

例えば、専門用語を追加すれば良くなるとか語と語の共起関係を追加すれば修正可能であるとかいう原因コードを付け修正にまわす。

4）原因コード点数別に分類する。

このようにして翻訳システムの改良を系統的に行うことができる。さらに、問題点の原因コード別に集計し、統計をとり、品質向上の参考資料を得ることができる。

〔4〕機械翻訳システムの品質検査

実際に稼働している機械翻訳システム上で2万件強のテストデータで翻訳を行ってみた。数社の協力が得られた。これらについては個々に企業名を出して発表することはしないが、翻訳結果を印刷する中で、どの企業のものが良いか簡単に判断することができた。

英語の教員に内容を検討してもらいその結果も参考にした。

〔5〕おわりに

機械翻訳システム（英⇒日）のテストデータを作成する方法として、大量の英文データを単語数別に分析し、整理し、単語数の少ないものから順次テストするという方法を考えついた。これにより機械翻訳システム（英⇒日）の品質の向上をはかることができる。

〔6〕参考文献

（1）安田賀計　らくらく使えるビジネス文書1230文例
CD－ROM　日本経済新聞社

（2）田久保浩平，橋本光憲　英文ビジネスレター
文例大辞典　CD－ROM　日本経済新聞社
15,000文

（3）社 日本電子工業振興協会
「自然言語処理システムの動向に関する調査報告書」
平成9年4月

〔7〕データについて

英文データは日本電子化辞書㈱の英文コーパスを利用した。

日本電子化辞書㈱のプロジェクトに参加した企業がEDR英文コーパスをあまり利用していない。翻訳システムにもっと活用してほしいものである。