機械翻訳用のテストデータ

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

〔1〕はじめに

これまでに、筆者は機械翻訳システムの性能評価と品質向上について考えてきた。その結果コーパス・ベース機械翻訳用のデータを集めるためにはどのようにすればよいかを調べ、その中で幾つかの重要なことが分ったのでそれをまとめて述べる。

〔2〕機械翻訳用の知識データ

機械翻訳は1つの言語から他の言語への変換を行うものである。

ルネッサンス革命はギリシャ、ローマ文化をヨーロッパ各国言語へ変換することから始まった。これに独自の文化を加え、新しいルネッサンス文化が花開いたのである。

我々は、今までの紙に書かれた科学技術と文化（本の知識）を電子化することで新しい文化と科学技術（電子化された知識）を作りあげようとしている。デジタル革命、インフォメーション・テクノロジー革命（IT革命）を行っているのである。

このためには機械翻訳は重要であるし、今までの紙に書かれたものを電子化しなければならないのである。特に日本語と英語のバイリンガル・パラレル・コーパスを作らなければならない。これを基にして知識データを積み上げていかなければならない。

バイリンガル・パラレル・コーパスはただ単に一般分野ばかりでなく、専門分野のバイリンガル・パラレル・コーパスが重要である。

これらのデータを基にバイリンガルのフレーズ、専門用語、複合語、文型パターンを大量に抽出しなければならない。コーパスから各種の知識を抽出し、体系化し、電子化しなければならないのである。

〔3〕機械翻訳用のテスト・データを集めるにあたって

機械翻訳用のテスト・データを集めるにあたって、次の点が重要である。

(1) 大量のデータを集める。

(2) バイリンガル又はマルチリンガルのパラレル・コーパスを集める。

(3) 機械翻訳システムに訳文自動評価システムを追加する。人手による検査を省くため。

(4) バイリンガル・パラレル・コーパスから簡単に訳文生成ルールの自動・半自動生成ツールを作成する。

(5) 専門分野別のバイリンガル・パラレル・コーパスを作成する。

(6) テスト・データは非公開にするか、もし公開にするならば毎年数回変える必要がある。

以上(1)〜(6)のことが重要である。次にそれらについて述べる。

(1) 大量のデータを集める。

大量とは約10万文を1つの目安として考え、毎年この程度異なった文をテストデータとして使用し、改良に用いなければならない。

これらのデータを持っているところは出版社や各種協会のようなところである。

又、大学で学生にタッチタイプの練習として入力したデータを集めるのも、一つの方法である。

もう一つの方法として、インターネット上の新聞記事からバイリンガル・テキストを大量に集めることを行っている。これは政治、経済、社会面の記事で、長文のデータが集めることができる。

毎年、最低10万文程度のデータはテスト用に使用したいものである。

(2) バイリンガル・パラレル・コーパス

モノリンガル・コーパスでは訳文の評価をする際に、正しい訳文がなければ判断に迷うし、正しい訳文をただちに作り出せる専門家を雇わなければ評価が出来ない。これは訳文の評価に費用と時間がかかる。

テスト文を評価に用いる際には、英文では単語数別ファイルにしてテストすることが望ましい。日本語文では、文字数別のファイルにして評価すべきである。これは文の複雑度の程度を文の長さで判断する基準になる。

英語を中心としたバイリンガル・パラレル・コーパスは他言語（日本語、英語以外）の言語の機械翻訳システムの作成にも有用である。

(3) 機械翻訳システムに訳文自動評価システムを追加すべきである。

単語数別ファイルを作り機械翻訳システムを評価する方法は、問題点の抽出が容易で興味ある方法である。

しかし、翻訳結果の評点付けは人間の作業であり、大変労力のかかる作業である。この作業を自動的に行うように機械翻訳システムに自動評価システムを組込んでおくことが重要である。最終的な判断は人間の作業であるが、ある程度のところまでは機械的に可能である。これにより翻訳結果の大まかなグループ分けが可能である。このようにして作業の迅速化が図られる。

例えば

1）構文解析で曖昧さが減らすことができない。
2）未知語が出現した。
3）意味解析のパターンが無い。
4）専門用語が無いため合成訳を作成した。
5）並列表現の解析がうまく行えなかった。

等、機械処理システムで誤る場所は幾つかある。これら

Test Data for Machine Translation System
Yasuhito Tanaka
Hyogo University

の問題点について重み付けを行い、評点を付けるのも一つの方法である。自動的評価システムを作らなければならない。

このような自動評価システムは人間の評価値と少し異なるかもしれないが、改良しなければならない文を大量の訳文の中から早く見つけ出すためには有効な手段ある。このようにして問題点を修正し、ほぼ完了した時点で人間の精密な検査を行い、さらなる改良に使用すべきである。

(4) バイリンガル・パラレル・コーパスから簡単に訳文を生成するルールの自動・半自動生成ツールを作成する。

機械翻訳システムは構文解析を中心としたルール・ベースのシステムから例文又は例文を一部変形したものを用いた例文ベースの機械翻訳システムに移ろうとしている。このための解析・生成用のルールの半自動生成ツールを作成すべきである。この点については、どのようにしたものが有効であるか筆者は研究中である。このようなルールは数１０万に達するものを作成することが必要である。

(5) 専門分野別のバイリンガル・パラレル・コーパスを作成する。

我々は学生達とバイリンガル・パラレル・コーパスを大量に集めてきた。しかし、このコーパスには次のような特徴を持つことが分かった。

1) 短い文が多い。(平均６単語)

2) 人称代名詞（I, you, He, She, They, We, …）等が主語となる文がほとんどである。

3) 日常的に使われる文が多い。

しかし、新聞や機械、電気、電子、情報、医学、等の専門分野の文はこのようなものとは異なっている。

それ故、分野別のバイリンガル・パラレル・コーパスを作成しなければならない。

各分野別に１０万～２０万文程度は集めるべきである。

(6) テスト・データの公開・非公開

テスト・データは非公開で行うか、公開するかは大きな問題である。しかし、テスト結果を公表するにあたってはどうしてそのような結果になったのかを示さなければならない。そのためたとえ非公開にしても少しずつ内容がわかってしまう。

公開するならば毎回テストごとに内容を変えてゆかねばならない。前回のテスト結果との比較については少し考慮しなければならない。

〔４〕バイリンガル・パラレル・コーパスを集めるにあたって

我々は学生とこのようなコーパスを作ってきた。

１年目	５万文	1997年（準備の年）
２年目	合計１２万文	1998年
３年目	合計１６万文	1999年
４年目	合計２０万文	2000年

４年間で２０万文　日本語・英語のパラレル・コーパスを作成することができた。これは機械翻訳製作者のテストデータとして使用した。非常に有効であることが分かった。しかし、専門分野の文が少ない、長文が少ないことも分かった。

学生達とこのように文を集めたが、実際にはこの倍程度の文を入力した。文の重複（訳文も含めて）が多いことが分かった。

これは次の３つの理由に起因している。

1) 学生が入力作業を怠り、他人のデータを流用する。

2) 同一文章が使われている。

例　i) good morning.　おはようございます。

ii) What time is it now?　今何時ですか。

基礎的な慣用表現は多くのテキスト中に出現している。

3) 同一の出版社の本の中には出版社の持っているテキスト例文を用いるため同一文が現れる。同一の著者の出版物にもこのようなことがある。

それ故、ただ単純に文を集めればコーパスができるものではない。

〔５〕より良い機械翻訳システムに向けて

より良い機械翻訳システムに改良するために次のようなことを提案する。

(1) 標準的な大量のテストコーパスを作成する。

毎年１０万文程度の標準的なコーパスを提示し、これらを最低限完全に翻訳することを課題とする。この例文をどのように誰が提供するかが課題である。

(2) 機械翻訳システムが多くの利用者に使われることを、第三者による評価が必要である。

多くの利用者に安心して機械翻訳システムを利用してもらうためには、第三者による評価付けが必要である。

これはただ単純に評点付けばかりでなく、それがどのような分野のどの程度の文で評価されているかということが、分かるものでなければならない。評点と同時に価格、特徴なども分かるものが必要であるし、定期的にこれら内容が更新されるものでなければならない。

日本ではインターネット上のホームページで機械翻訳システムを評価しているものがある。これは１つの参考資料になる。

hht:／www.bekkoame.ne.jp／～ot3／

の　Green and White　のページを参照するのも一つの方法である。

銀行の財務内容によってランク付けがなされるものと似ている。

〔６〕おわりに

我々は今後一層機械翻訳システムの品質向上のために研究を続けてゆかねばならない。機械翻訳システムは研究つくされた分野ではない。まだまだ未解決な分野が多い。少しずつの改良が大量に積み重なると、前のものとは本質的に異なったように改良される。機械翻訳の改良は全体的に前より悪くなる改良はない。悪くなるような改良は中止すればよいのである。

〔７〕参考文献

(1) 田中康仁　機械翻訳システムの評価と改善
情報処理学会　自然言語処理研究会　133－1
1～6pp　1999年9月

(2) ㈱ 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