機械翻訳用のテスト・データについて

田　中　康　仁

愛知淑徳大学

0）はじめに

機械翻訳用のソフトウエアが各メーカーやソフトウェア会社から発売されている。しかしこれらソフトウェアの品質は今一歩というところがある。機械翻訳用ソフトウェアの品質を向上させるためにはどのような方法があるか考えてみた。

ここではパラレルコーパスを作り、テスト方法を自動化、又は、半自動化することにより、何が悪いか、何を追加すればよいかを知る方法を考えてみる。

1）どのようにして検査するか？

機械翻訳用ソフトウエアを使ってみると思うような結果が出力されないと思っている人々は多いと思う。そこで内部情報を公開してほしいと言っても企業側は企業のノウハウの公開は出来ないと言ってしまう。

しかし、我々は日常、多くの工業製品を使っている。自動車、洗躍機、日本語ワープロ等を使っている。これからの製品は故障しないことはない。各メーカーは故障情報を代理店、販売店、消費者、消費者団体、テスト専門の企業、雑誌社…等から集めている。問題が明確なものから、漠然とした問題まで巾広く情報を集めている。もし、集めて新しい商品を作られなければ競争から脱落してゆく。

このような工業製品と同じように機械翻訳用ソフトウエアについても悪い点を明確に示す方法と大量のテスト・データを準備すべきであろう。

このためには次のような方針を立てる。

1．大量のテスト・データを集める。

2．テストのための同一の基準データを作成する。年々この量を増加させる。

例えば日本語↔英語の対になってパラレルコーパスを準備する。

3．明確な証拠をもとに、評価付けを行う。

4．結果を公表する。

このような方針で集められたデータを基にして消費者団体を動かし、改善をせまれば機械翻訳ソフトウエアは良くなるはずである。もし、このような競争をさけて、脱落すれば、その会社の製品は売れなくなるのである。我々は内部情報を知らなくても良いのである。

2）どのようなテスト・データが良いか？

テスト・データとしてパラレル・コーパスが良いと述べたが、それはどんなものか、形式を例で示してみよう。

入力データ→ 工業製品 →出力データ

図1　工業製品のテスト

例1

［000001］←Seq No.

［英　語］［Would you mind writing down the name and addres of the school？］

［日本語］［学校の名前と番地を書いて下さいますか］

［注　釈］［Would you mind writing it down. は丁寧で上品な言い方である。普通は　Please write it down.］

［出　典］［秋山登志之著　ポケット英語決まり文句辞典　P 76］

英語の解説書には英語、日本語の対訳と同時にちょっとした注釈がある。又、出典 page No.も重要である。

さらに、次のようなキーワードの項目を設定することも重要である。

例2

［000002］←Seq No.

［英　語］［Where were you goofing off？］
［日本語］［どこでサボってたんだ］
［注　釈］［職場を離れる場合］
［英語キーワード］［goofing off］
［日本語キーワード］［サボる］
［出　典］［南雲堂　実用英語決まり文句辞書　佐藤誠司　小池直巳　P 72］

テスト・データは一文で意味が明確になるものとし、文相互間の関係は少ないものを選ぶことにする。このような目的の資料、例文集は数多く出版されているので集めることは簡単である。しかし、著作権のことを考えておかなければならない。入力作業はなるべく単純作業とし、その後編集してまとめる。

3）試行と結果

このような構想に基づいて2社の機械翻訳システムでテストした。日本語、英語の対応した文を約100文ばかり翻訳した。2社にはテスト目的は言わず修正しないことを条件とした。A社、B社の人々はそれぞれ資料を提供してくださった。しかし、この結果は1994年春の状況であるため、その後の更新版ではこのような結果は出なくなっているであろう。

A社の例（1）

［000003］
［英　文］［The large family system is not suited to the present social conditions of Japan.］
［日本語訳］［大家族制は日本の現在に合わぬ。］
［機械訳］［大きい家族制度は、日本の現状の社会コンディションに適していない。］

この出力効果から〝large family system〟は〝大家族制〟として複合語の登録をしたほうが良いことがわかる。又、同様に〝social condition〟は〝社会状況〟として複合語登録をしたほうが良いことも判る。このようにすれば機械訳は〝大家族制度は、日本の現在の社会状況に適していない。〟となり良い出力結果に変わることも予測できる。

A社の例（2）

［000004］
［英　文］［Blue looks good on you］
［日本語訳］［紺がきみによく合う］
［機械訳］［青は、あなたの上で良いように見える。］

この結果は〝look〜good on you〟が〝〜によく合う〟という訳を機械が持っていないために機械訳が出力されているのである。
動詞句の抽出にも役立つ。

B社の例（1）

［000005］
［英　文］［They have brought me a box made of wrong measurements］
［日本語訳］［寸法の合わぬ箱をもってきた］
［機械出力］［それらは、私に間違っている測定で作られていた箱を持って来た。］

これは、〝made of wrong measurements〟を〝寸方の合わぬ〟と訳せばよいものをこれができないために少し悪い訳文となったのである。このように改善されれば〝それらは、私に寸法の合わぬ箱をもってきた。〟となり良い出力結果になることが判る。

B社の例（2）

［000006］
［英　文］［I undid the package and found the contents damaged］
［日本語訳］［その小包を開けたら中身はいたんでいた］
［機械出力］［私はパッケージを取消し、内容が破損されたのを見つけた。］

これは〝undo the package〟を〝小包を開く〟という意味に訳さないために起った誤りである。共起情報の不足のために起った誤りである。
この文の問題点は英文、日本語訳、機械出力を比較することにより簡単にわかる。高校卒業、程度の能力があり、少し訓練を受ければ改善の方法がわかる。
この分析結果を［追加注釈］［……］としてテストファイルに蓄積してゆけば同じようなシステムの分析に役立つし、検査の人員を減らし、能率を上げることができる。

4）自動照合システムの概要（機械翻訳用）

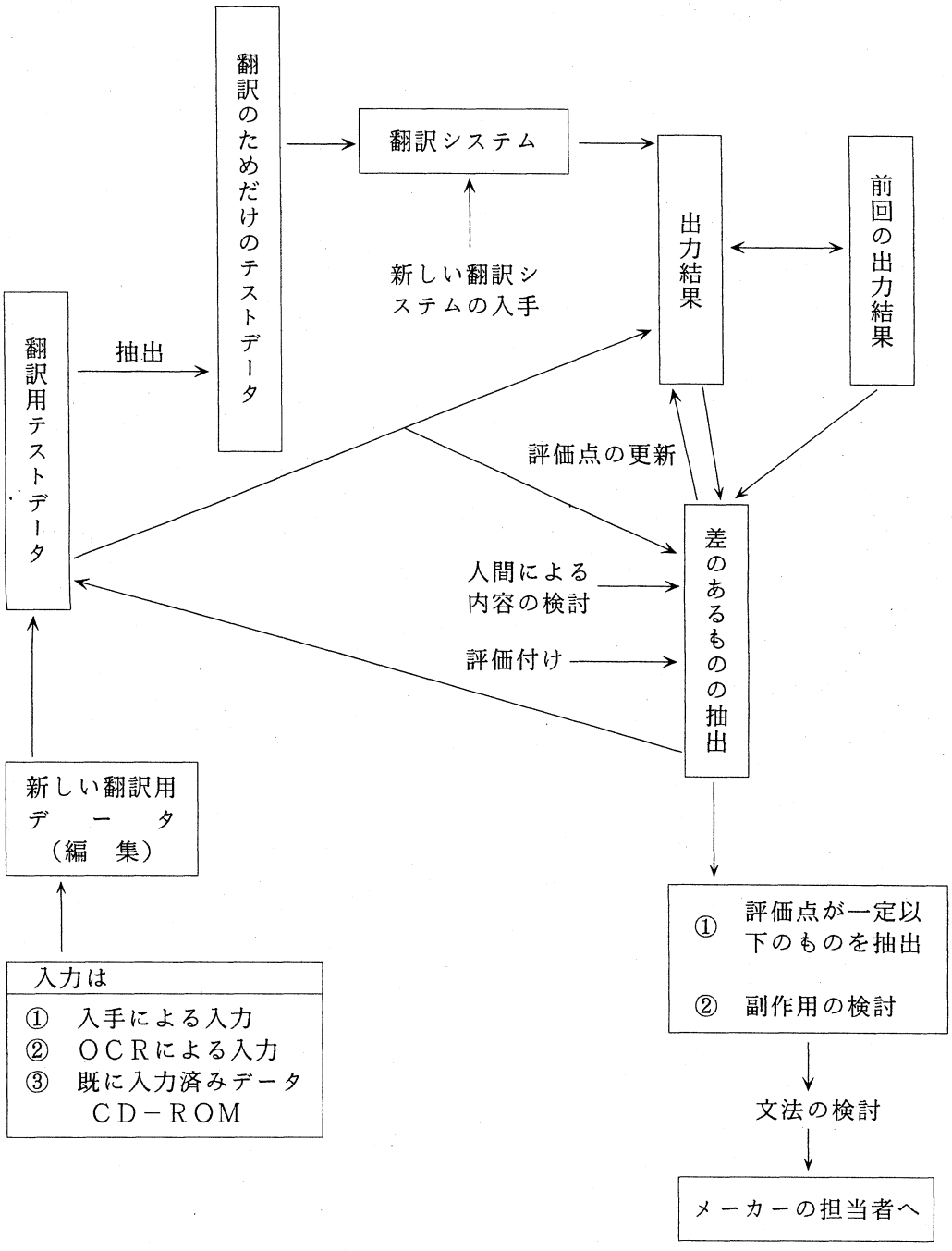

図2　自動照合システム（機械翻訳用）

図２のような考えを自動照合について持っている。機械翻訳システムも個々の言葉の使用条件をこまかく規定し、入力してゆく時代になってきた。
これは大変な作業であるが着実に行ってゆけば可能である。
一年間に５万件のテスト・データを作るとすれば、５年間で25万件の例文から知識データが抽出できる。このようなグループが５ケ所あると125万件の例文になる。着実にデータを集めることを考えれば困難なことではない。
まさに「量的拡大は質的変化をもたらす」ことになる。

５）パラレルコーパスの利用方法

パラレル・コーパスの一つの利用方法として機械翻訳用テスト・データとしての例をあげたが、これらコーパスはその他種々の利用方法が考えられる。

① 用例ベースの機械翻訳の例文として使用できる。
② 日・英の構文解折用文型パターンの作成資料となる。
③ 教育用教材として使用する。
④ 自然言語処理の各種の基礎データとして役立つ。
⑤ 辞書作成用の資料となる。
⑥ その他

①～⑤等が利用方法として考えられる。今後も利用方法を検討してゆきたい。

６）おわりに

昭和58年４月県立姫路短期大学に着任したころから「語と語の関係」という考え方で語の共起関係を利用し自然言語の曖昧性解消のために努力し、仮名漢字変換の同音異義語の解決に向けてデータを集めてきた。この結果を利用し各社の日本語ワープロは品質が向上したようである。
５年ないし、10年努力しパラレス・コーラスを集めれば機械翻訳の性能向上にも寄与できるものと確信している。多くの人が考えれば照合システムももっと自動化できる。
脳力のない機械翻訳システムに智恵をつめこむことができる。

７）参考文献

① ㈳日本電子工業振興協会
機械翻訳システムの実用化に関する調査研究　1993年３月
② ㈳日本電子工業振興協会
機械翻訳の開発と実用に関する実態調査
1989年７月
③ 秋山登志之　ポケット英語きまり文句辞書
南雲堂　1987年７月
④ 小池直已、佐藤誠司　実用英語きまり文句辞典　南雲堂新書　1989年11月

８）資料と謝辞

この研究に使用したデータの一部は電子技術総合研究所知能情報部自然言語研究室　元吉文男室長の許可を得て同研究室で作成した講談社和英辞書（㈱講談社刊）の磁気ファイルを使用した。
（このデータ作成は前記研究所員、現、岐阜大学工学部池田教授等によって行われた。）記して感謝の意を表す。

９）筆者からの連絡

平成７年４月１日より次へ移ります。
兵庫大学　経済情報学部教授　田中康仁
〒675-01　兵庫県加古川市平岡町新在家
2301番
TEL　0794（24）0052
FAX　0794（26）2365