7M-4

機械翻訳の用例データベースについて

田中康仁

兵庫大学

E−mail:yasuhito@humans−kc. hyogo−dai. ac. jp

〔０〕はじめに

用例を基にした機械翻訳システムの提案が行われて久しい、この間、用例の類似度についての計算式やその問題点については多くの人々によって議論されてきた。しかし、データについては議論されていない。ここではデータについて考え、分野別用例データベースを機械翻訳システムに登録し利用することを提案する。

〔１〕用例データについて

用例を基にした機械翻訳向けのパラレル・コーパスの作成を考えると次の問題点があげられる。

ｉ）どのような分野を考えればよいか？
ii）どのようにして用例を集めるか？
iii）どれだけの量を集めれば、どの程度の成果が得られるか？
iv）どれだけの費用がかかるか？
v）容量について?

ｉ）〜V）について考える。

１）専門分野別バイリンガル・コーパス

機械翻訳のカタログを見ると次のような分野がある。

１、情報処理	１４、生物
２、電気・電子	１５、〔医学〕生化学
３、物理・原子力	１６、〔医学〕薬学
４、機械	１７、〔医学〕解剖学
５、工業化学	１８、〔医学〕疾患症状
６、プラント	１９、〔医学〕精神医学
７、土木建築	２０、〔医学〕医療機器
８、金属	２１、金融・経済
９、地学・天文	２２、法律
１０、輸送	２３、ビジネス
１１、自動車	２４、人名・地名
１２、軍事	２５、環境
１３、農林水産	

（富士通、ＡＴＬＡＳＶ６英日・日英翻訳ソフトカタログより）

このほかに一般的分野が考えられる。このようなデータは次のようなものを考えればよい。

一般的分野としては

・英語検定試験用データや参考資料
・ＴＯＥＩＣ（Test of English for International Communication）の試験問題や参考資料
・中学１〜３年の教科書、高等学校１〜３年の英文

このようなものを集めればよいと思う。これらの資料はグレード別のデータである。これを集めテストする意味があるし、性能を調べるために都合がよい。

２）どのようにして用例を集めるか？

専門分野別のバイリンガルで書かれているカタログ、説明文、解説文を含むマニュアル等を多量に集め機械可読ファイルを作成する。著作権についても考慮する。

３）データ量について

第１段階として約１０万文程度を集めることを目標とする。これには３人で１年数ヵ月で入力することができる。

・１年の労働日数を２００日とする。
・１日で１人のバイリンガル・データの入力文数は約２００文とする。
・２人が作業し、１人がデータ・チェックを行う。
・１年に５分野を行うとして５〜６年程度で完了する。

これを行うには各分野の人に協力と分野別データのための著作権の問題解決が重要である。

４）費用について

機械翻訳システムが定着してきている。今後、この機械翻訳システムがなくなることはない。市場原理が働きコスト・パフォーマンスの悪いもの、改良に向けての資金力のないものは市場から無くなっていく。そのためにいかに改良費用を作り出していくかが重要な課題である。着実な改良が望まれる。次の式が成り立つ。

改定版の売上−改定版作成の総費用＝利益

改定版作成の総費用のうち何%をデータ作成費用とすることができるかがポイントである。

５）容量について

用例データを３００文程度入力するとＴｅｘｔ形式で約３０ＫＢ〜４０ＫＢ程度である。

１つの分野について１０万文程度入力すると約１０ＭＢ〜１３ＭＢ程度である。２５分野では約２５０ＭＢ〜３２５ＭＢ程度である。さらにデータの圧縮技術を用いれば約１０分の１程度にはなるので特に問題はない。さらにパーソナルコンピュータの容量も近年目覚ましい技術進展があり、容量の拡大が容易になってきた。いまやＧＢ単位での容量が使用できる。

〔２〕テンプレートを利用した翻訳例

従来の構文解析や結合価文法だけでは、こなれた文章を作ることはできない。用例データを少し加工しテンプレートが容易に作成できる。ここではテンプレートを用いればこれら

Example Data for Machine Translation System
Yasuhito Tanaka
Hyogo University

の問題点を解決できる。

英→日の翻訳の場合

「She has no heart for this work.」

〔N1〕has no heart for 〔N2〕

→〔N1〕は〔N2〕が気に入らない。

このようなテンプレートがあると次のように訳文が生成される。

「彼女はこの仕事が気に入らない。」

日→英の翻訳の場合

「若者で携帯電話を持っている人が増えた。」

〔N1〕で〔N2〕を持っている人が増えた。

→〔N1〕who has 〔N2〕has increased.

Young person who has the cellularphone has increased.

この2つの例は富士通の翻訳例のカタログから引用した。

このようにうまくテンプレートが適合すれば良い訳を作成することができる。

第1段階として非常によく使われる文はそのままとする。例えば

1）What time is it, now? ←→今何時ですか。

2）good morning ←→ おはよう

慣用的に使われる文はテーブル・ルックアップによって、ただちに出力される。

日本語、英語の対になった文は翻訳者が例文参照する場合にも有効である。

第2段階として日本語と英語の対になった文からエディターを使ってステレオタイプの対訳文を作り出す。

例えば

He went to school. ←→彼は学校へ行った。

→〔N1 (He)〕went to school. ←→〔N1（彼）〕は学校へ行った。

→〔N1 (hum)〕go to school.←→〔(N1 (hum)〕は学校へ行く。)

少しずつ抽象化してステレオ・タイプの対訳を作り出す。

ステレオタイプの対訳パターンを分類して整理し、同一のものは省いてゆく。

〔N1〕だけでなく〔N2〕も追加してゆく。

〔N1 (hum)〕go to school.←→〔N1 (hum)〕は学校へ行く。

〔N1 (hum)〕go to 〔N2(cons)〕←→〔N1 (hum)〕は〔N2 (cons)〕へ行く。

このような手作業の中からコンピュータによる自動的作成方法を考える。

最終段階のステレオタイプ・パターンばかりでなく、途中段階のパターンも機械翻訳システムにうめこむと有効である。

ここで次のような問題点がある。

1）テンプレートを機械的に作成するには、バイリンガルパラレルコーパスをどのように変形させればよいのであろうか。

2）何個ぐらいの〔N1〕、〔N2〕、〔N3〕……を一文の中に作ればよいか。

3）何を〔N1〕、〔N2〕、〔N3〕とするか。

4）作ったテンプレートがどのくらいの量があればどのくらいの適合率があるか。

これらの問題を解決しなければならない。

〔3〕効果の測定

一般文のバイリンガル・用例を約12万文対用いることで約20％ほどの性能向上が得られている。これはA社の実績である。ただ単純に用例だけの効果ではない面があるかもしれないが、効果の大きいことが分かる。

また、用例を用いて結合価文法の文型を強化したり、訳文の生成のためのテンプレートも増強することができる。分野別の用例データが有効なのは次のような理由だからである。

各専門分野では特別な言いまわしが普通に使われているし、動詞も特別なものが使われている。

例をあげる。

例1）英：The edges have been machined true and square to each other.

日：縁を、正しくかつ互いに直角に削った。

例2）英：A tap is used for internal threading.

日：タップは雌ネジ切りに使われている。

分野別の用例の増強についても、専門分野別に調査しなければならない。

〔4〕機械翻訳システムの利用面から

機械翻訳システムは一括翻訳ばかりでなく翻訳者が一文ずつ機械と対話しながら機械翻訳システムを用いて翻訳する場合がある。この時、似た例文を探したり、専門用語を検索したり、似たような言いまわし方を探す事が多い。このためにも専門分野別用例データ・ベースが必要である。

専門分野の訳語選択のセンスワードを作成することにも役立つ。

〔5〕おわりに

ここでは、機械翻訳システムの用例データベースを各分野ごとに設定する事を提案し、その効果等についても調べてみた。このようにして機械翻訳システムが品質向上され、使いやすくなることを期待するものである。

〔6〕参考文献

1）Freda Steurs : Machine aided Termbank Construction. Development of an Integrated and Learning System for Semi－Automatic Annotation. A Technological tool for Translation memories. TKE'99 Terminology and knowled Engineering TERMENT.

2）Yasuhito Tanaka, Kenji Kita JCKE Multilingual Corpus of Major Asian Languages. TKE'99 Terminology and knowled Engineering TERMENT.

3）富士通 ATLASV 6英日・日英翻訳ソフトカタログ

4）野沢義延 機械を説明する英語 工業調査会 1998．7

5）野沢義延 続機械を説明する英語 工業調査会 1997．6

6）郑保山、刘群、張祥 机器翻訳模板的建立原則和方法計算語言学文集 PP299－3004，1999年11月 清華大学出版社