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）

ＩＢＭの「翻訳の王様＋」では基本辞書に加え、次のような専門分野の辞書を作っている。

1）インターネット	5）アート
2）エンターテイメント	6）スポーツ
3）ビジネス	7）科学
4）政治	

となっている。

これは翻訳システムの販売分野によって専門分野辞書を作っているのである。

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）どのようにして専門分野の優先順位を考えるか？

どのようにして専門分野の優先順位を考えるかは重要な問題である。

これは考える人の立場によって、その方法が異なってくる。ここでは幾つかの方法を示す。

（ｉ）国家がどのような科学技術又は文化を導入しようとしているか。又は外国へ科学技術や文化を送り出そうとしているか。

これは国家としてどのような文献の翻訳物が多いかを調べればよい。

私の推定では、情報、コンピュータサイエンス、通信の分野、生物特に遺伝子工学の分野、精密工学、自動車等が考えられる。

（ii）機械翻訳システムを販売する人々からの要求

機械翻訳システムを販売する人々の要求に合わせて優先順位を考えればよい。つまり、市場の要求に合わせればよいのである。市場の要求に合わせ専門分野を決めればよいのである。

（iii）専門分野についての協力者

色々な専門分野を考えても、その分野の専門家が協力してくれなければ何もならないものである。そこで、専門分野の協力者がみつかった分野から順次実行する。そしてなるべくその分野の人々に協力してもらい、入力もその分野の人々に行ってもらうのがよい方法である。

４）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。

Example Data for Machine Translation System

Yasuhito Tanaka
Hyogo University

・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。

5）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。

6）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〕今後のデータ入力について

専門分野別に１０万文の例文体（日本語←→英語）を入力し、効果を実際に測定し、さらに今後どのようにするか考える。

（i）例文の選択が良かったか？

（ii）例文の量が充分であるか、今後例文の増加でどの程度の性能の向上が期待できるか等を考えてみる。そしてさらなる例文入力を考える。

〔５〕どのようにして文型パターンを作成するか。

自動的又は半自動的方法で文型パターンを作るかが重大な問題である。これを解決するには幾つかの方法がある。

（ｉ）バイリンガルのパラレル・コーパスを用いて手作業で順次作成する。この方法は確実であるが、時間がかかる。

（ii）バイリンガルのパラレル・コーパスをそれぞれ形態素解析、統語解析を行い、それにより文形パターンを作る。

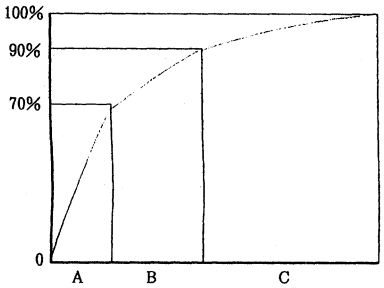
しかし、形態素解析、統語解析に曖昧さが発生するし、構文木が一意に決定しずらい面もある。

このような問題点が解決できれば大変良い方法である。

（iii）ＡＢＣ分析による単語の分析より半自動に作成を行う。

バイリンガルの英文コーパスの単語を集め、その頻度付のリストを作る。これを基礎にして頻度の高いものから順次並べる。

累積頻度が７０％までのものをＡランク、７０～９０％のものをＢランクとし、９０％～１００％のものをＣランクとする。



Ｃランクの単語について自動的に〔 〕を入れて文型パターンを自動的に作り出す。Ｃランクの単語が連続するならば一つにまとめる。

数詞は自動的に〔 〕に入れる。そして文型パターンを作る。

I am a 〔 professor 〕←→私は教授です。

その後、手作業で日本語の教授に〔 〕を入れる。このようにすれば一部自動化されるので、かなり作業が単純化される。

このプログラムの開発、Ａランク、Ｂランクの単語の選択も自動的に行えるので、大変良い方法である。

Ａランク、Ｂランクの単語に加え、情報検索や文書の索引の自動作成の際に使用する不要語リスト（Stop List）を使用することも考えられる。

これらを合わせ、再検索するのも一つの重要な点である。

この方法については、ＥＤＲの英文コーパスを用いて実験を行っている。

〔６〕パターン抽出の実験

６－１　ウオール・ストリートの単語分析

ウオール・ストリート・ジャーナル１年分新聞記事データから単語の抽出と分析をした。ここでいう単語は文字列の意味である。（例えば単数形と複数形は別単語となっている。）

単語総数		298,737単語
延単語数		26,848,148単語
ランクＡ単語	943	18,794,335
ランクＢ単語	4,496	5,368,767
ランクＣ単語	298,737	2,685,046

以上のような結果が得られた。

６－２　ＷＷＷの単語分析

英文で作成されたＷＷＷを取り込み、この単語を分析した。ここでいう単語は文字列の意味である。

単語総数		248,434単語
延単語数		9,757,136単語
ランクＡ単語	1,989	6,829,911
ランクＢ単語	12,883	1,958,458
ランクＣ単語	233,562	968,767

以上のような結果が得られた。

６－３　各分析単語の特徴

ウオール・ストリートのランクＡ，Ｂの単語は5,439単語、ＷＷＷのランクＡ，Ｂの単語は14,872単語であった。ＷＷＷのランクＡ，Ｂの単語はウオール・ストリートの約３倍程であった。

これはＷＷＷの総単語が少なかったのとＷＷＷが広い範囲から集めているので、このようになったのである。

６－４　文型パターンの自動抽出実験

ウオール・ストリート・ジャーナルのランクＡ，Ｂとストップ・ワードでの実験結果はＷＳｊで示す。

ＷＷＷのランクＡ，Ｂとストップ・ワードでの実験結果はＷＷＷで示す。

各ランクＡ，Ｂ，STOPWORD 以外の単語は〔 〕で囲む、数詞は無条件に〔 〕で囲む、また〔 〕が連続する時は一つに纏めた。

EDR文中単語10の結果

〔〕の数	WSj	WWW
0	867	3,242
1	2,579	3,446
2	3,252	1,848
3	1,919	515
4	474	61
5	23	2
合 計	9,114	9,114

EDRコーパスの全部の文について

WSj	%	WWW	%
11,006	8.75	39,297	31.24
30,333	24.11	43,129	34.28
34,954	27.78	26,615	21.16
26,301	20.91	11,826	9.40
14,843	11.80	3,788	3.01
6,194	4.92	947	0.75
1,783	1.42	169	0.13
324	0.26	25	0.02
63	0.05	7	0.01
7	0.00	0	0.00
1	0.00	0	0.00
125,803	100.00	125,803	100.00

この結果からみると、WSjの結果の方が良いこと

がわかる。単語数は5,500単語におさえ、それにストップ・ワードを追加し、6,000単語が良いことがわかる。しかも、なるべく機能語（冠詞、前置詞、接続詞、be動詞、省略語、代名詞、疑問詞等）と基本動詞とその変化形を多くするようにすべきであり、名詞をなるべく少なくすれば良いということが分かった。〔 〕が一つも含まない文があるが、これは基礎的語で出来た文である。これらの中には慣用表現が多く含まれていた。

この方法の利点を述べる。

1） プログラムが簡単であり、テーブルに入れる単語を調整することでプログラムの働きがよく分かる。
2） 統計やnグラムで使用頻度の高い部分文字列を抽出するより、プログラムの働きが分かるので使いがってが良い。
3） 少数のデータに対しても使える。
4） テーブルを基礎語にすることにより文の選別ができ、むつかしい文とそうでない文への選別ができる。これは外国語の教育やレベル別の指導用文が得られる。
5） テーブルに入れる基礎語が機械的に抽出できる。この方式で文型を自動的抽出は無理である。一部分は手作業で修正しなければならない。このためには、情報処理の専門家と言語の専門家との協同作業が重要である。これは、英文での研究と実験であるが、日本等にも応用できる。少しずつ量を増やし、テーブルを修正し、量を少し増やし、テーブルを修正するという作業が重要である。

6－5 一部分手作業で抽出した文型の処理

手作業で抽出した文は自動的に文型パターンを展開する。例えば三つの〔 〕がある文を考えると、次のような七つの文型の可能性が自動的に作られる。

——〔 A 〕——〔 B 〕—〔 C 〕—。
1）—〔 A 〕————————。
2）————〔 B 〕————。
3）————————〔 C 〕—。
4）—〔 A 〕——〔 B 〕————。
5）—〔 A 〕————〔 C 〕—。
6）————〔 B 〕—〔 C 〕——。
7）——〔 A 〕—〔 B 〕—〔 C 〕——。

このように自動的に展開し、有効な文型に印を付け、選択すればよりよい文型抽出ができる。

〔 〕の数と展開される文型パターンの数を調べると次のようになる。

〔 〕の数	展開できる文型数
1	1
2	3
3	7
4	14
5	34

〔 〕の数が6ケ以上のものは手作業で展開するか、それらについては非常に少数（1.4%程度）であるので、例外的と考え何も展開しない。

専門分野別のバイリンガル・コーパスでは、専門分野で非常に多く使われる語を含む文型パターンも簡単に作り出すことができる。

〔 7 〕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。

〔 8 〕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、文型パターンの自動抽出システムや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。

〔 9 〕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） 徳永健伸 「情報検索と言語処理」第2章 PP20～21 東京大学出版会 1999
7） 羽鳥洋美 宮平知博 辞書の自動切り替え機能を考慮した翻訳辞書 情報処理学会 第61回（平成12年後期）全国大会 1T－1 2000年10月
8） 斉藤健太郎 池原 悟 村上仁一 大規模コーパスからの重文、複文の統語構造の自動抽出 情報処理学会 第61回（平成12年後期）全国大会 3T－3 2000年10月

注）この研究は、２０００年９月ポーランドで行われた"Translation and meaning"の国際会議で一部発表し、その後ポーランド日本情報工科大学で講演した際の討論を基に追加し、内容を深めたものである。パターンの自動作成のことについては、その時に考えたものである。