6M−02

例文を用いた機械翻訳システム
── 例文の一致度について ──

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

〔１〕はじめに

例文機械翻訳システムに必要な例文をどのように収集するかについて方法を議論し、手法を確立した。

ここでは、それらデータがどの程度有効であるかについて検討した結果を示す。

〔２〕例文機械翻訳システムの有効性について

例文機械翻訳システムのために日常会話文を中心に約２１万２千件の日本語文と英文の対になったデータを集めた。

この作業は約５年間の期間がかかった。またデータの中には形式の不整合やデータの誤り等が混在するため、修正作業が大変であった。今も修正作業を行っている。

このデータの名称をＡファイルとする。

このデータとは別に２００１年度前期に学生達がタッチタイプの練習として入力したデータを整理した。

(1) 入力したデータ総数　　３４，１４７件

これらデータの中にはコピー等を行い提出したもの、同一の本や資料を使用したために発生する重複がみうけられる。そこでこれらを削除した。

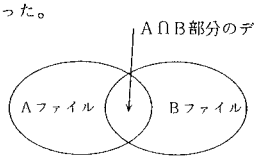
(2) 重複を削除したデータ件数　　２７，４８０件

６，６６７件の重複がみつけられた。

このデータには日本語、英語に重複はない。このデータの名称をＢファイルとする。

一部の学生の中には前年度受講した学生もいるため、Ａファイルのデータの一部とＢファイルのデータに同一データを提出した可能性もある。また同一の高等学校や同一のテキストや練習問題帳を使用している者もいる。このため、ＡファイルとＢファイルを完全に独立なファイルとみなすことはできない。このような問題点は存在するが、次の実験を行ってみた。

どの程度の重複がＡファイルとＢファイルで見られるか実験を行った。



ＡとＢの共通データは８，１４７件であった。

用例の有効性は　８１４７／２７４８０＝０．２９６

約３０％のデータが既に集めた２１万件のファイルと一致した。

Example Base Machine Translation System
Yasuhito Tanaka
Hyogo University

この実験結果には検討しなければならない課題がいくつかあるが、おおよそ２０％～３０％の一致度が考えられる。これはある同一の分野に限った場合のことである。

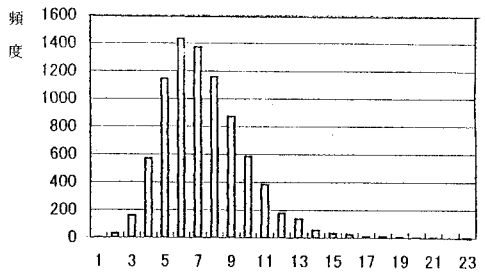
これまでに集めたＡファイルに原子力のデータを照合させても前述のような実験結果は得られない。

これを単語数別の図とグラフに示すと次のようになった。

1	2	3	4	5	6	7	8	9	10
3	30	158	569	1,145	1,432	1,369	1,157	872	584

11	12	13	14	15	16	17	18	19	20
382	178	134	55	31	24	8	8	3	2

21	22	23	計
2	0	1	8,147



このグラフ等から判断して１３単語以上のデータは、同一の学生のデータ又は同一教材によるデータと考えられる。このような実験を数回行い、それらをまとめ、ある一定以上の頻度のものをまとめれば正しい統計データが得られると確信している。

例文を用いた機械翻訳システムでは、１２単語までは例文として用いるべきである。それ以上のものについては、次の条件を満たす以外は使用しなくてもよいと考える。

１．ディスク容量に充分な余裕が有るとき。
２．例文ファイルを余分に持つことにより、例文の検索速度が遅くならないとき。
３．長い例文を利用することが可能なとき。
　人と機械が調和して翻訳しにあたることができる手段（即ちインターフェイス）があること

これらを満たす時には長文もデータファイルとして保存すべきである。

長文の例文の中には次のような性質の文も数多く存在する。

１）文法書等に用いられている代表的な例文
２）ことわざなどの文
３）有名人の発言や小説等の中の名文

これらは積極的に入れるべきである。

機械翻訳システムを開発しているソフトウェア会社は用例翻訳のシステムを組込んでいるようであるが、どれだけの用例があるのか、どのような分野か、どの程度の効力があるのか示してほしい。

〔３〕分野別のデータ・ベース

我々は日常生活で用いられる文を中心に集めてきたが、特定の分野別の日本語と英語が対になった用例データ・ベースを集めるべきである。

それではどのような分野を考えるべきであろうか。これは利用者がどのような分野を望んでいるかという分析が必要である。

富士通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）

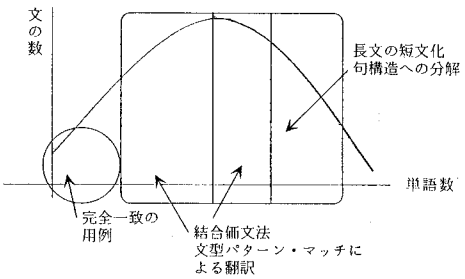
ＩＢＭのカタログには次のような分野がある。

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	

〔４〕用例データ・ベースの利用

用例データ・ベースは単純に文を照合させ一致するものを利用するという方法もあるが、文の一部を変数化し、文型パターンとして翻訳に利用することもできる。

また、長文については文型パターンの例文としても利用できるが、もっと単純な構造に変形させるための例文として利用すべきであろう。それ等については今後の研究で示したい。



〔５〕今後の課題

次のようなことを考えて発展させたい。

１）このコーパスを拡大したい。

２）分野別パラレル・コーパスへの発展

３）長文の解析を考えたい。

４）長文の解析のための、日本語、英語の対になった句の蓄積を行わなければならない。

５）その他

これらの課題を研究し、機械翻訳、自然言語処理の発展につとめたい。

〔６〕おわりに

用例文の有効性がどの程度あるか調べてみた。約２０～３０％程度あることがわかった。しかし、これはある特定の分野の用例を２０万件程度集めて成り立つことであることも重要である。

〔７〕参考文献

(1) 田中康仁　機械翻訳システムの今後について
情報処理学会　自然言語処理　１３７－２
２０００年６月

(2) 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　１３３－１
１９９９．９

(3) Yasuhito Tanaka,　Kenji Kita
JCKE Multilingual Corpus of Major Asia Languages
TKE' ９９ Terminology and Knowjedge Engineering
１９９９．６

〔補足追加〕

複数の協力者のタッチ・タイプの練習とコーパス・データ作成という目的をかかげて行う作業では、一部の人々は他人のデータの全部または一部分をコピーして提出するということが行われている。

これは次のような方法で発見することができる。

１）提出されたデータを１つにまとめる。英語順、日本語順に分類する。

２）１）で作成したファイルの中で日本語、英語が完全に一致する重複データを抽出する。

３）個人毎のデータが重複データを何％含んでいるか調べる。

４）重複データの含まれている％の高いものを調べる。２０％以上のものは要注意である。

１）～４）の作業を短期間に行うことが重要である。

このような方法は確立したが実際には行わなかった。不正を行って、タッチ・タイプの練習をしなければ、それなりに苦労するだけである。