対訳付慣用表現の収集について（その2）

4L－3
田中康仁
兵庫大学
Email:yasuhito@humans-kc.hyogo-dai.ac-jp

[0] はじめに

機械翻訳の基礎的知識（Basic Knowledge）としては単語、複合語、専門用語、概念、概念と概念の関係、慣用表現等が考えられる。個々に研究を進めなければならないが、私の考えでは複合語と慣用表現の対訳付データの大量の収集と研究、体系化がなかなか進んでいないと考える。これは機械翻訳の応用分野がある特殊な分野に限られているためで、広くて一般的な分野に移れば、この分野は重要な問題になると考えている。数量としても数万から数十万の慣用表現データが必要である。

また高級な知識とは:「水は摂氏℃0度で凍る。」とか「アメリカは原爆を日本の広島と長崎に落とした。」「その結果、数十万の人々を殺した。」というようなものである。高級な知識、又は基礎的な知識の中間に位置するようなものも考えてゆかねばならない。例えば諺等がそれである。

[1] 複合語について

複合語は我々の身のまわりにある語で2語以上の自立語の組合さったものであれば良いと考える。これらは曖昧さを大幅に減らすことができる。自立語としては動詞、名詞、副詞、形容詞、形容動詞をさす。2語以上の複合語といっても必ずしも曖昧さを減らすものではない例もある。

例 white house: 白い家、アメリカ大統領府

しかし、英語で2語以上の自立語を集めることは機械翻訳に役立つのである。

これら複合語の特殊な分野としては、次のようなものがある。

1) 地名:日本国内の地名、世界の地名、都市名
川、河、山、湾、湖、海、……等
2) 機関名:日本の国、地方公共機関の名前、役人の地位、団体、協会、学会等の機関名
3) 会社名, 組織名、局名、部名、課名、役職名
4) 氏名: 姓、名（男性名、女性名）
5) 物品名: 生活用品、家具、薬品名、病名……
6) 日常語、標語
7) その他

以上等が考えられる。

これらのものが備えられている市販の機械翻訳システムは少ない。

[2] 慣用表現について

慣用表現については情報処理学会第52回（平成8年前期）全国大会4B-9で述べたが、さらに研究成果を述べる。この中で、

1) 集められた慣用表現についてどのように標準化するか？
2) 慣用表現と訳しわけ

以上については問題があるということのみを述べ詳細は省いた。そこで、これらについて調べる。

2－1) 慣用表現の標準化

慣用表現は文の中から抽出し次のような標準化を行わなければならない。

1) 動詞の過去形は原形にする。
was able to→be able to
is ,was , were……→be
2) 所有形について
in his wayhome→in one's wayhome
かえり道
3) 再帰代名詞→oneself
dried himself off →dry oneself off
体をかわす
4) 複数形は単数形にする
soft contact lenses→soft contact lense
5) Theはaにする。
しかし、特殊なものはそのままにする。
例 The Japanese a Japanese
国民全体 個人

How do we extract Idiomatic Knowledge?? (NO2)
Yasuhito Tanaka
Hyogo University

6）短縮形は短縮でないものにする。

don't → do not

I've → I have

7）語をまたがるものについては［A］［B］等の記号を入れる。

launch［A］in［B］←→［A］を［B］に送り出す。

8）動詞＋ing→動詞

driving wasps off→drive wasp off

はちを追い払う

9）その他

標準化の一般的規則を述べたもので色々の例外がある。個別に検討しなければならない。

2－2）慣用表現と訳しわけ

慣用表現には色々な訳語が付いている。

例 best season: 最も良い時、良い時期

しかし、実際には次のように使われている。

The cherry is best season.

桜は満開である。

このため、慣用表現の訳語としては代表的な訳語を採用している。

しかし、実際には代表的な訳（複数の場合有り）がどの程度使われているか、別訳（表記上のちよっとした場合と、異なった意味の場合がある。）がどのように使われているか、その条件は何か等を調べてゆかなければならない。

2－3）どのように集め入力するか、入力形式について述べる。

1）No:ナンバリング

2）英文慣用句

3）和訳（複数可能）

4）書籍記号

5）頁番号:該当する頁番号

これらは逆スラッシュ（＼）をもちいて区切りを行う。次のように入力した。

例 ＼1＼05515

＼2＼be unable to ［A］

＼3＼［A］することができない

＼4＼F

＼5＼12

＼1＼05502

＼2＼call at

＼3＼訪ねる

＼4＼F

＼5＼10

2－3）入力データ件数

8冊の本から慣用表現を抽出し入力、整理した。1つの英語の慣用表現にどの程度の訳語が付しているか調べてみると次のような結果になった。

1種類	6，170件
2種類	2，888件
3種類	478件
4種類	88件
5種類	21件
6種類	6件
合計	9，653件

これらについてはさらに内容を整理、検討している。

［3］今後の課題

慣用表現については次のことを調べなければならない。

1．対訳付コーパスの収集、どの程度のコーパスが有用か？

2．標準化された慣用表現を用いて対訳コーパスからの例文の自動抽出

3．集められた慣用表現（対訳付）が、他の方法で集められた慣用表現（マルコフモデルを用いて、英文だけの中から抽出したもの）と、どの程度一致するか。種類延べ件数ではどの程度であるか。

以上のようなことを考えてさらに発展させたい。

参考文献

1）田中康仁、吉田 将 慣用表現について
ー収集と整理ー情報学基礎5-1 情報処理学会 1987.6

2）田中康仁 対訳付慣用表現の収集について 情報処理学会 第52回（平成8年前期）全国大会4B-9 1996.3

3）田中康仁 吉田 将 概念辞書の作成 自然言語処理75-12 情報処理学会 1990.1