言語的類似性を利用する日韓音声翻訳の検討

白 京姫　白井 諭　山本和英　坂本 仁
ATR 音声言語通信研究所
E-mail: {kpaik, shirai, yamamoto, msakamo}@slt.atr.co.jp

1 背景

機械翻訳は、一般に言語解析、言語変換、言語生成の3つの処理を直列に配置することにより構成される。各処理の精度向上が翻訳精度につながるが、辞書の語彙数のほか、変換規則の多寡が与える影響が大きい。このため、両言語に通じた人が両言語の相違点を変換規則として精細に記述し、これを蓄積していくことにより、精度向上が図られてきた。しかし、翻訳精度の向上が優先的に進められ、翻訳知識の開発コストと翻訳品質の関係に関する議論はあまり行なわれていない。開発コストの削減は機械翻訳を開発するに当たっては重要な要件であると考えられる。

どの言語対を翻訳対象とする場合でも単語や句単位の変換辞書は必須である。一方、文法が大きく異なる言語対では多くの変換規則が必要になるのに対し、文法が類似していれば変換規則の記述量は少なくて済むと考えられる [成田 96]。その極端な場合として、変換規則がゼロの状態ではどうなるかが定量的に示されているわけではない。本研究では、文法が類似している場合に、変換規則なしでどの程度の翻訳が可能か、また、どのような変換規則を加えればどれだけ翻訳精度が向上するかを段階的に解明していくこととしたい[1]。もう1つの重要な課題である変換辞書の構築については別稿に譲る。

文法的な類似の基準として、本稿では文献 [Gre78] に示されている分類のうち、語順に関する次の3つの類型情報に着目した。

(1) 名詞の構文上の働きを規定する語は前置 (pr) か後置 (po) か

(2) 形容詞 (A) と名詞 (N) の語順

(3) 主語 (S)、目的語 (O)、述語 (V) の語順

[1]音声翻訳への適用を考えると、翻訳品質が高いに越したことはないが、いくらかでも翻訳できれば役立つ場合もあると考えられる。

表 1: 語順に着目した言語の分類例

po	AN	SOV	Bashkir Bengali Buriat Burmese Gujarati Hangarian Huichol **Japanese** Kannada Konkow **Korean** Kurku Mongolian Ossetic Panjabi Piro Quechua Telugu Turkish Uzbek Vogul Yakut
		SVO	Finnish Guaarani Ojibwa
	NA	SOV	Basque Chitimacha etc.
pr	AN	SOV	Amharic
		SVO	Chinese English Russian etc.
		VSO	Chontal Squaamish
		VOS	Tagalog
	NA	SOV	Persian Tajik etc.
		SVO	French Thai Vietnamese etc.
		VOS	Malagasy
		VSO	Arabic Hebrew Samoan etc.

これらを、(1) を pr と po、(2) を AN と NA、(3) を SOV や SVO などと表記すると、表1のようになる[2]。

本研究では文法的に類似する言語間の翻訳を対象として、変換規則の量と翻訳精度の関係を検討する。その一環として、本稿では、po-AN-SOV に該当する言語相互の翻訳を取り上げる。具体的には日韓翻訳を対象とするが、原則として日本語または韓国語に依存しない簡便な翻訳手法を検討する。

2 日韓両言語の類似点と相違点

日・韓両言語は様々な面でよく似た特徴を示す。その類似点のうち、主なものを以下に挙げる。

(1) 日本語と韓国語は多くの漢語を使う。

(2) 語順が概ね同じである。

(3) 敬語の体系が似ている。

[2]参考のため po-AN-SOV は全言語列挙した。また Korean は文献 [Gre78] には記述がないので我々が追加した。

(4) 語感に対する発想が似ている。

日本語と韓国語は多くの漢語を使うが、その七割が同じであると言われている [渡辺 81]。また日韓両言語はどちらも中国から語彙 (および漢字) を借用し、自国語の語彙に取り入れてきたため、日本語で音読みする漢字語は多くの場合にそれを韓国語として読むだけで翻訳可能である。例えば、下記のように、日本語から韓国語に翻訳を行なう際には、韓国語で該当する訳語に置き換えることで、多くの場合翻訳ができる。

日・韓両言語- の- 類似点- と- 翻訳- との- 関係
일・한양언어- 의- 유사점- 과- 번역- 과의- 관계
il・hanyangeoneo-uy-yusajeom-kwa-beonyeok-kwauy-kwankei

また、上記の例は名詞句だけの例であるが、語順が同じであることと, 日本語に該当する韓国語がそのまま置き換えられていることが分かる。文のレベルでも両言語は SOV(Subject-Object-Verb) 系であり、助詞などは名詞の後に来るという面で類似している。

そして、両言語とも類似した敬語の体系を持つ。ただし、日本人が身内の人に対しては敬語を使わないのに対して、韓国人は自分より年上の人に対しては必ず敬語をつかう。他の言語の敬語システムに比べると、非常に似ている体系を使う。日本語・韓国語は同じ漢字文化を共有する言語であるがゆえに、その文化でなければ理解できない、言葉に対する共有できる感覚がある。敬語の使い方も儒教文化の共有というところから生まれたと思われる。「まっかな嘘」は直訳しても韓国語で通用する。しかし、日本語で「顔が広い」という表現は、韓国語では「足が広い」のように表現しなければならない。このように似たような発想からかえって間違いが生じるとも言われるが、一般的には翻訳が容易であると言える。

今までの機械翻訳もこれらの類似点を利用してきたと思われるが、さらにそれを最大限活かしてどこまで翻訳できるかを本研究では追求し検討していきたい。勿論、機械翻訳処理の難題である両言語間の相違点も注目すべきである [Kim98]。本研究では追って相違点と翻訳品質の定量的な関係を検討することとしたい。

3 翻訳処理の構成と試行実験

本節では、日韓翻訳の簡単な翻訳手法を提案し、試行実験の結果を述べる。

3.1 処理の構成

手法が現実的なものであるためには、必要とするデータは容易に入手できるものでなければならない。ただし、変換辞書の実現は極めて大きな課題であるが、詳細は別稿に譲り [白井 01][3]、ここでは変換辞書の存在を前提として、翻訳対象言語の類似性を前提とした処理の構成を検討する。

機械翻訳は、一般に、原言語解析、言語変換、目的言語生成により行なわれる。原言語解析の最初に行なわれる形態素解析では、原言語表現に含まれる単語を見出す役割を担っている。このため、単語辞書を参照して単語の可能性を総当たり的に探索して単語グラフを作成し、別途作成しておいた単語の連接の可否を判定するための統計情報を用いて、単語分割の妥当性を判定する方法が多く行なわれている。これを類似言語間の翻訳に応用すると、次のような構成が考えられる。

1. 原言語表現を変換辞書と照合することにより、単語候補とその訳語を単語グラフの形で取り出す
2. 単語グラフの適当なリンクごとに目的言語コーパスを検索し、出現度数を妥当性として集計する

ただし、日韓翻訳では、上記の構成のほかに次の 2 点が必要となる。

日本語の動詞や形容詞等は活用するので、この語形変化に対応することが必要である。日本語の活用は正規文法で記述できることが知られているが、本稿では茶筌 [松本 00] で代用する。変換辞書との照合では単語候補が得られなかった部分を対象として、それが茶筌の単語開始位置と一致していれば、1. の結果に追加する。

韓国語では日本語と異なりわかち書きが必要となる。2. で目的言語コーパスを検索する際、単語候補を結合した場合と単語候補の間に空白を狭んだ場合の 2 種類の検索を行ない、出現数が多い方を選択する。

以上の方法は、変換辞書を別にすれば、単言語コーパスは対訳コーパスに比べはるかに入手が容易であることから、容易に実現できると考えられる。なお、機械翻訳の原言語解析では、単語レベル、構文レベル、意味レベル等での多義性解消も目的の 1 つにあげられるが、ここでは文法的類似を前提として、この問題には立ち入らない。

次節では、上記の方法による日韓翻訳の試行実験について述べる。日韓変換辞書の規模は約 35,000 語で

[3] 英語を介することにより生成する方法を検討中である。

あり[4]、日本語1語に対し1個以上の韓国語表現が対応付づられている。また、韓国語コーパスは約20,000文[5]である。

3.2 翻訳例

日本語と韓国語をあまり知らない人に日韓翻訳をやらせたとしよう。そして、その人は日本語と韓国語がかなり似っているという事実だけを知っているとする。この場合、訳をするために、与えられた日本語文に含まれた語を一つずつ、または色々な連鎖で引いてみる方法がある。ここでの翻訳はまさにそのレベルの話である。即ち、変換ルールにまったく頼らず、語と語の対訳辞書だけ利用してどのような翻訳ができるのかを以下に示す。

下記はATRの日韓旅行会話集から取り出した文である。Jは日本語の入力文を表す。Kは日本語の入力文を予め翻訳した訳文である。Kを正訳文として見做して、対訳辞書だけで翻訳された文と比較してみよう。

- J：一番安いシングルはおいくらですか。
- K：가장 싼 싱글은 얼마입니까?

まず、次のように対訳辞書を引き始める。

(1) "一"からはじめ、語になりうるものを全部引き出す。:"일 -il"、"하나 -hana"、 "갈음 -katteum"、"제일 -jeil"ののような韓国語の対訳が選ばれる。

(2) "一番"を取り、同じようにそれに当たる韓国語の対訳を探す。

(3) "安い"を"安"から引き始める。

(4) "シングル"は "シ"、"シン "、"シング"のような組み合わせは日本語にはないので、"シングル"にまとめて韓国語を探した結果、"싱글 -singeul"と訳すことができる。

この結果次のような辞書引き結果が得られる（[付録] 参照）。

```
1. "一" "一番" 普通名詞      "가장" adv
                             "맨"   adv
          "一"   普通名詞    "일"   cn
                             "하나" cn
                             "갈음" cn
                             "제일" cn
2. "番" "番" 普通名詞 "번" cn
3. "安" "安い" 形容詞 "싸"   adj
                      "저렴" adj
4. "い"
5. "シ" "シングル" 普通名詞"싱글" cn
6. "ン"
7. "グ"
8. "ル"
9. "は" "は" "는" top
10. "お" "おい" 感動詞 "어이" misc
         "お"   感動詞 "어"   intj
11. "い" "いくら" 普通名詞 "얼마" cn
         "いく"   本動詞    "가" verb
12. "く"
13. "ら"
14. "で" "ですか"      "ㅂ니까"     aux
         "です" 本動詞  "입니다"     misc
                        "말입니다"   misc
         "で" 格助詞    "에서"       post
                        "로"         post
                        "으로"       post
         "で" 普通名詞  "그래서"     func
                        "그러니까"   func
         "で" 接続詞    "근데"       conj
         "で" PAT       "그래서"     conj
         "で" PAT       "그리고"     conj
15. "す"
16. "か" "か"  終助詞  "하는가"      post
                       "는가"       post
         "か"  普通名詞 "모기"      cn
         "か"   PAT   "ㄴ지"       misc
         "か"   PAT   "ㅂ니까"     aux
         "?"                       symbl
```

この例では茶筌の解析結果を参照したことにより追加された訳語はない。

3.3 問題点

韓国語コーパスにより単語連接の可否を判定すると次の結果が得られる。ただし、"/"はその前後の単語がコーパス中に見出せなかったため、空白を入れるべきかどうかの判定ができなかった箇所を示す。

```
1. 가장 저렴 / 싱글 / 는 어 / 얼마입니다하는가.
2. 가장 저렴 / 싱글 / 는 어 / 얼마입니다는가 /?
3. 가장 저렴 / 싱글 / 는 어 / 얼마입니다모기 /?
```

[4]文献 [古瀬99] の実験に使用された日韓変換辞書を拡張した。拡張に当たっては文献 [山本00] に基づいて韓国語の品詞を付与した。

[5]文献 [Fur94] のコーパスの一部に韓国語訳を付与した。

前節のように対訳辞書だけでここまで訳すことができた。通常様々な変換ルールを用いて訳せても、これより少し良い程度である。勿論、上に示したようにこの翻訳方法だけでは助詞の問題や動詞の語尾変換の問題は全然できてないだろうと予測した。しかし、それを除くと、ここでの問題は対訳辞書を改善することによって解決できる。例えば、”安い”が「形容詞 (”저렴” adj)」となっている。しかし、形容詞の処理が辞書上の形容名詞になっていて、本来形容詞につく語尾がついてないため、「”저렴/싱글” (安さ/シングル)」のような不自然な訳になっている。そこを「저렴한」(”安い”)を辞書に登録することにより、文末の語尾変化以外は良い訳になる。実際、初めから変換ルールを構築し翻訳に取り組もうとするとかなりの開発コストを要する。しかし、辞書作成が終わった段階であれば、本研究で取り上げたような翻訳機構を実現するのは極めて容易である。

4 まとめ

本研究では、類似言語間の翻訳において、翻訳知識の開発コストと翻訳品質の関係を定量的に示すことを目的として開始した。特に、(1) 名詞に後置詞が付加される、(2) 形容詞が名詞に先行する、(3) 主語－目的語－動詞の語順をとる、という特徴を持つ言語間の翻訳を対象とし、第1ステップとして、変換規則を使わない日韓翻訳実験を開始した。今後は、この方法により受容可能な翻訳がどれくらい得られるか翻訳失敗の原因がどのような言語現象によるか、等を定量的に把握していく予定である。

参考文献

[Fur94] FURUSE, O., Y, S., TAKEZAWA, T., and URATANI, N.: Bilingual corpus for speech translation, In *AAAI-94 Workshop on Integration of Natural Language and Speech Processing*, pp. 84–91 (1994).

[Gre78] GREENBERG, J. H.: *Universals of Human Language*, Vol. 4, Chapter Syntax, Stanford University Press (1978).

[Kim98] 金泰完, 崔杞鮮：日韓機械翻訳システムの現状分析及び開発への提言, 自然言語処理, Vol. 5, No. 4, pp. 127–149 (1998).

[古瀬99] 古瀬蔵, 山本和英, 山田節夫：構成素境界解析を用いた多言語話し言葉翻訳, 自然言語処理, Vol. 6, No. 5, pp. 63–91 (1999).

[山本00] 山本和英：計算機処理のための韓国語言語体系と形態素処理, 自然言語処理, Vol. 7, No. 4, pp. 25–62 (2000).

[松本00] 松本裕治, 北内啓, 山下達雄, 平野吉隆, 松田寛, 浅原正幸：日本語形態素解析システム「茶筌」, version 2.2.1 使用説明書 (2000), URL=http://chasen.aist-nara.ac.jp.

[成田96] 成田一：言語類型と機械翻訳, 情報処理学会 研究報告 96-NL-114-21, 情報処理学会 (1996).

[渡辺81] 渡辺吉鎔, 鈴木孝夫：朝鮮語のすすめ - 日本語の視点から -, 講談社 (1981).

[白井01] 白井諭, 山本和英, 白京姫：対訳辞書の生成のための英訳辞書の照合, 電子情報通信学会 技術研究報告 NLC (2001).

[付録]