対訳コーパスを利用した日英変換の分析

塚脇　幸代
s-tsuka{at}dream.ocn.ne.jp

概要

対訳コーパスから日英 2 言語間の「差異」を発見し、日英変換における課題を提示した。コーパスはアラインメントにより生成されており、スコア上位の 500 対を対象とした。抽出用に品詞タグをつけ、日英変換の抽出に利用した。日本語の品詞を手掛かりに、英語でどのような品詞に変換されているかを観察した。結果、アラインメントのスコアが高い対の中から、いくつかの差異が観察された。

はじめに

対訳コーパスを利用する目的として、外国語学習者が目的言語の表現を獲得するための教材や、外国語学習者の中間言語(Interlanguage)を分析するツールとして使うことが考えられる。本稿ではこうした利用法のほかに、言語分析のための利用を提案する。アラインメントにより対訳を生成する場合には、「何が同じであるか」が着目されるが、異なる言語の変換において、「何が異なっているのか」を分析することにより、日英変換の課題を発見することができる。また、2 言語間の違いと、各言語の個別言語としての特徴を見出すこともできると考える。まず、品詞タグをつけた対訳コーパスから、品詞単位での差異を観察する。また、差異が認められた日本語が、既存の翻訳システムでどのような訳出になるかを見る。そこから、変換処理の課題となるべき現象を見つけ出す。

1. 方法

日英変換のお手本として、「日英新聞記事対応付けデータ(JENAAD)」(Utiyama and Isahara, 2003)の一対一文対応ファイルに収められた 150,000 対のうち、タグ付けの終了したスコア上位 500 対を対象として、日英語の変換分析を行う。対訳コーパスの日本語文は ChaSen で形態素解析を行った後、対訳コーパス用のタグを付与した。英語文はスペースを単語区切りとみなして分割し、日本語文同様に対訳コーパス用のタグを付与した。[1]日本語文英語文ともに、対訳として参照しやすいように、「形容動詞語幹」＋「に」を副詞(ADV)としてまとめたり、句動詞を一単位にまとめる処理を人手で施した上で、日本語文および英語文を Microsoft® Access® 2010[2]データベースにリンクし、日本語側から「出現形」「見出し語」「品詞タグ」等を指定して、対となる英語文を検索できるようにした。図1は日本語の形容詞(ADJ)及び副詞(ADV)を指定して対訳となる英語文を検索した結果である。共通のテキスト ID を持つとして日英文をそれぞれタグとともに表示している。

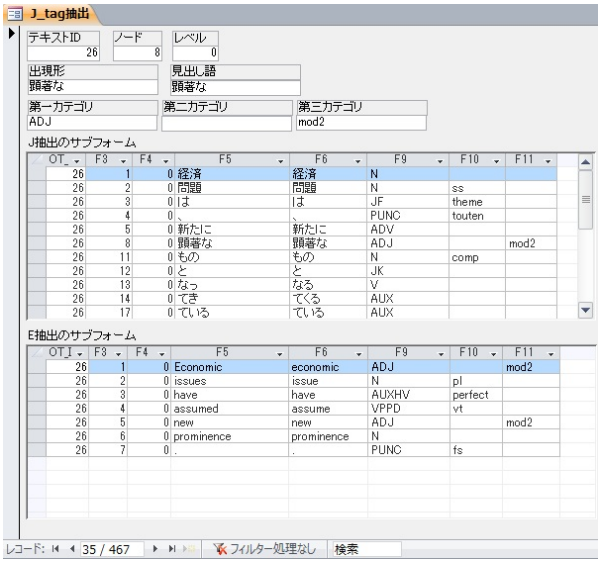

図1: 対訳文抽出結果

検索対象の形態素はノード番号で表示されているので、日本語文のノードにある単語が、英文でどのように訳出されているのを目視確認する。日本語の品詞と異なる品詞で訳出されているものを、日英間の構造上の差異として取り出した。今回は日本語の品詞が、(1)形容詞(ADJ)および副詞(ADV)、(2)動詞(V)であるものを中心に観察を行った。

2. 結果と考察

日本語と英語の品詞構成に差異があった対を以下に例示する。一行目はテキスト ID、日本語、日本語品詞タグ、英語、英語品詞タグである。二行目に日本語文、三行目に英語文を斜体字で示す。4行目と5行目には、現在

[1] 参考文献[3]の品詞タグを使用した。

[2] © 2010 Microsoft Corporation. All rights reserved.

利用可能な Web 翻訳サービスの出力結果を示す。[3]翻訳には 2 つの Web 翻訳サービスを利用した。Microsoft 社が提供する Bing Translator と Google 社が提供する Google 翻訳である。Bing Translator の出力には BT、Google 翻訳の出力には GT を、それぞれ行頭に示して区別する。

日本語の形容詞(ADJ)および副詞(ADV)からの変換では、ADJ→ADV への変換、および ADV→ADJ への変換が見られた。またこの変換が起こっている訳対においては、変換後の文構造にも変化が見られる。テキスト ID487 では、主語＋述語の SV 構造から、phrase へ変換されている。一方、テキスト ID472 では、ADJ の修飾先であった*損害*が、*を与える*とともに動詞 *damage* に吸収され、消滅している。

487, 激しく, ADV, fast-changing, ADJ
国際情勢が激しく変化する中で、日本の平和と繁栄を維持していく上で、外交・安保政策も大事だ。
Amid the fast-changing international situation, diplomatic and security policies are important in maintaining Japan's peace and prosperity.
BT: To maintain Japan's peace and prosperity among the international situation varies wildly, on foreign and security policies is important.
GT: In the rapidly changing international situation, on the go to keep the peace and prosperity of Japan, it is also important foreign and security policies.

472, 明白な, ADJ, clearly, ADV
決定の中で、「原告の訴えは、外交関係をめぐる連邦政府の行動に明白な損害を与える」などと述べている。
They stated in their conclusions that the plaintiffs' demands would clearly damage the federal government's activities relating to diplomatic relations.
BT: "Plaintiffs will damage obvious diplomatic relationships between Federal Government action", and said in the decision.
GT: I said, in the decision, "plaintiff's claim is obvious damage on the behavior of the federal government over the diplomatic relations" and so on.

テキスト ID16 のように、ADV の係り先が動詞から名詞に変更される例もある。

16, 最も, ADV, most, DET-qt[4]
自由貿易で最も利益を得ているのは日本ではないか。
It is Japan that has received the most profit by free trade.
BT: Or is not in Japan are getting the profit most in the free trade.
GT: Is not Japan have benefited the most free trade.

日本語の動詞(V)からは、形容詞(ADJ)への変換がみられる。テキスト ID489 は状態性の動詞から形容詞へ、486 は否定形から形容詞への変換である。

489, 値する, V, be_worthy, AUXBE_ADJ
真剣な検討に値しよう。
It is worthy of serious discussion.
BT: Worth serious consideration.
GT: Would deserve serious consideration.

486, 避ける_られる_ない, V_AUX_AUX, be_inevitable, AUXBE_ADJ
改革に伴う「痛み」は避けられない。
The pains that accompany reform are inevitable.
BT: "Pain" caused by the reform is inevitable.
GT: With the reform of "pain" is inevitable.

他品詞に変換される例としては、名詞(N)や、前置詞(PREP)への変換もある。

492, 発展する, V, growth, N
中国市場が世界で唯一、発展し続けるかどうかも問題だ。
Another issue is whether China will continue to be the only growth market in the global economy.
BT: Whether China continues to evolve only in the world too!
GT: Matter only, even if the Chinese market whether to continue development in the world.

493, 包囲する, V, against, PREP
世界の平和を脅かす国際テロ組織を包囲する広範な国際社会の共同戦線が、急速に構築されつつある。
The global community is rapidly building a united front against the international terrorism network that is threatening world peace.
BT: While the United Front of the broad international community surrounding the international terrorist threat to world peace, developed rapidly.
GT: A broad united front of the international community that surrounds the international terrorist organizations that threaten the peace of the world, is being built at a rapid pace.

また、テキスト ID490 は一単語が複数の単語に変換されている。

490, 有効活用する, V, make_effective_use, V_ADJ_N
日本企業は生産効率の追求は熱心だが人材や資金を有効活用する戦略を欠いている。
Japanese firms are zealous in their pursuit of improved production efficiency but lack strategies to make effective use of human resources and funds.
BT: Lacks the strategic pursuit of productivity's keen Japan companies exploit the financial and human resources.
GT: Lacks a strategy to make effective use of human resources and funds Japanese companies are eager pursuit of production efficiency.

変換後の品詞が変わることにより、文単位や句単位で日英間の構造が変化している。テキスト ID 485 は形容詞文から There is 構文へ、491 は前掲の 487 とは逆に、phrase から sentence へ変換されている。

485, 強い, ADJ, strong, ADJ-mod2[5]

[3] 2013 年 1 月現在の出力である。
[4] DET-qt は数量限定詞を意味する。
[5] ADJ-mod2 は限定用法の形容詞を意味する。

国内には、このように軍事力増強を続ける中国への援助継続に、批判的な声が強い。
There is strong criticism at home against continuing aid to a China that keeps building up its military might
BT: Aid continued to China keeps military buildup in this way, critical voices strong in Japan.
GT: China to continue aid to continue the military buildup in this way, in Japan, a critical voice is strong.

491, に対する, JKEQ, pose_to, V_PREP
台湾に対する中国の脅威は増している。
The threat China poses to Taiwan is increasing.
BT: Taiwan against Chinese threats are increasing.
GT: China's threat to Taiwan is increasing.

利用した対訳コーパスにはアラインメントのスコアが付与されており、最も高い訳対では 1.710796165、最も低い訳対では 0.04347244008 というスコアがついている。下位の訳対には、以下のように日英語の対応が完全には一致しない対も含まれる。

<T
ID="19930310J1TYMAG1400080/19930311E1TDY06C00002 0/8" NM="1-1" SCORE="0.043478
13198">
<J>この上、借金を増やせば、二十一世紀の子孫が借金地獄になる」</J>
<E>Issuing deficit-financing government bonds is like paying household food bills with borrowed money. </E>
</T>

分析対象に用いた 500 対には、最も高いスコアを持つ訳対から、0.4837899487 のスコアを持つ訳対が含まれている。日本語と英語の対応が機械的に比較的取れやすい対であるといえる。そのような訳対の集合の中では、日英間の差異は比較的小さいとも予想されるが、実際には前節でみたような異なる品詞構成への変換も見受けられる。
また、このような変換を、機械翻訳システムがどのように処理しているのか、2 つの Web 翻訳サービスの出力を借りて示したが、必ずしも対訳コーパスの通りではなかった。原言語の文構造と品詞構成をそのまま目的言語の文構造と品詞構成に置き換えることが翻訳の基本であると仮定すると、前節に例示したような訳対は、機械翻訳にとっては、少し高度な変換作業であると言えそうである。もっとも、どのような処理と計算によって出力がなされ、どこで失敗しているのか、出力結果からは断定しかねる。当該箇所が訳出として不適切であるとき、他のさまざまな要因でそうなっていることもあるからである。しかし、そのことを差し引いても、形容詞(ADJ)→副詞（ADV）、副詞(ADV)→形容詞（ADV）またはその逆において、とくに修飾用法で変換が起きた時、処理が追いついていない印象がある。

3. 議論

ある言語から異なる言語への変換は、それが人間の手によるものであれ、機械によるものであれ、原言語の意味を損なうことなく目的の言語の形式に生成されることが求められる。しかし一つの意味に対し、生成可能な表現形式は、原言語および目的言語ともに複数の形が可能であり、どの形を選んでも構わない。原言語の文構造と構成要素がそのまま目的言語でも使用される変換が基本であるとするなら、文構造を変える、または句レベルで構造を変えるという処理については、
（1）その言語にとって自然な表現であるか。
（2）その言語の他の表現形式と比べて適切か。
（3）その表現が文脈にあっているか。
（4）その文体は特殊ではないか。
等々を考慮すべきである。たとえば、前掲のテキスト ID491 の「に対する」に対し、前置詞（PREP）を訳に用いるのは自然で無理のない範囲であろう。一方、テキスト ID500 の「する」は、そのまま訳出すると不自然になる。そこで意味を解釈して適切な訳出をすることになる。

500, する, V, be_considered, AUXBE_VPPD
企業の強制分割は「最後の手段」にしたい。
Forcible division of businesses should be considered a last resort.
BT: Forced division of companies want to "last resort".
GT: Companies want to make a forced break "last resort".

「する」などの漠然とした表現から解釈をするのは、機械にとって難しい作業であるが、人間にとっても、漠然とした言い回しは訳語に困ることがある。上の例では、「したい」とは特定の誰かの動作ではなく、主語を立てることができない。したがって対訳文では受動態が選択されている。
文脈をみなければならない場合、一訳対の範囲で判断がしづらいばかりでなく、何をつけたし、何を省略するべきかという問題に直面する。人によっても判断が異なると考えられ、ここで選択肢は大いに広がる。文体の話までふくめると、さらに選択肢は増える。また、あまりにも意訳過ぎると対訳としては落ち着きが悪い。そのなかから何を適切な訳とし、どのような訳出を目指すかは、人間と機械に共通の課題であろう。それは同時に、翻訳の評価の問題でもある。目指すべき訳出があれば、おのずと評価も定まる。利用法を間違えなければ、対訳コーパスから指針を得ることができる。各言語には、それぞれその言語としての自然な表現があり、訳出の手本となる対訳コーパスにはなるべく自然な表現が収められているのが望ましい。

4. まとめ

対訳コーパスを利用して、日本語から英語への変換がどのようになされているか、とくに訳出文において品詞が異なる事例を見つけ出し、観察した。観察結果から、形容詞(ADJ)→副詞（ADV）への変換をはじめ、いくつかの品詞変換が行われている訳対を発見した。またそれらの訳対が現行の機械翻訳システム上ではどのように訳出されるかを確かめた。

おわりに

利用できる対訳コーパスは数が限られており、日英新聞記事対応付けデータ（JENAAD）は大変貴重な言語資源である。対訳コーパスを使って翻訳の妙味を引き出せないかと思い立ったのが、タグ付きデータを作成するきっかけであった。本稿で利用したタグ付きデータは2005 年に出来上がっていたが、先頭 500 から増やせていない。今後効率よい方法をみつけることができればすべての訳対について処理を行いたい。

参考文献

[1] Masao Utiyama and Hitoshi Isahara. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003, pp. 72--79.

[2] 形態素解析器 ChaSen http://chasen-legacy.sourceforge.jp/

[3] 塚脇幸代 (2005).対訳コーパスのためのタグ体系，言語処理学会第11回年次大会. P3-1.

[4] Bing Translator http://www.bing.com/translator/

[5] Google 翻訳 http://translate.google.co.jp/