言語処理学会 第2回年次大会発表論文集, pp.221–224, 1996

英日機械翻訳における速読支援のための日本語生成

九津見 毅　奥西 稔幸　佐田 いち子
シャープ株式会社 情報商品開発研究所
{kutsumi, okunishi, sata}@isl.nara.sharp.co.jp

1　はじめに

現在の英日機械翻訳システムの多くは、一文単位で解析し翻訳処理をするものであり、文より小さい句や節をまとまりとして解析・生成するようにはなっていない。だが、現状の翻訳システムでは、英語と日本語のように語順が大きく異なる言語間で、係り受けの複雑な入力文が入力されると、たとえ正しく構文解析・生成しても、論理的に間違ってはいないが読みにくい日本語文が出力されることが多い。

昨今、低価格の英日機械翻訳システムが普及し始めているが、その主な使われ方をみると、精訳を得るのが目的というよりも、英文読解を支援するシステムとしての利用が多くみられる。

そのような利用法を想定した、訳文を英文に添付して出すという表示方法の翻訳システムが出現しているが、そのような出力形態を採用するなら、それにふさわしい日本語を出すことが必要であると著者らは考える。いうなれば英語の語順に近い日本語であるが、日本語だけを単独で読んでも不自然でないことがもちろん望ましい。

このような目的のために著者らが実用化したのが、原文を適当な単位で切り分けて、それに対して訳を付ける翻訳システムである。

図 1: 翻訳エンジンのブロック図

2　システムの概要

図１は著者らが採用している英日機械翻訳システムの翻訳エンジンのブロック図である。このシステムをベースとして、原文を適当な単位（便宜上「フレーズ」とよぶ）で切り分け、そうして定めた各々のフレーズに対する訳（「フレーズ訳」とよぶ）を生成するモジュールを、このシステムの最終段階である「日本語生成」部に付加したものが、本稿で紹介する翻訳システムである。

たとえば、

> In addition to sequential files, the second file structure used by SHARP-BASIC is random access.

という英文を翻訳する場合、図１の「木変換」まで処理が進むと、図３に示すような木構造が得られる。さらに処理を進めて、図１の「日本語生成」の最終段階まで進むと、図２に示すように、訳文を構成する訳語が、各々の原文単語と訳文単語との対応情報を伴って得られる。このたび開発したシステムでは、フレーズ訳を生成するための特別な処理をこの段階で行う。

この特別な処理は、分割位置決定処理と、フレーズ訳生成処理そのものとの２段階に大別される。分割位置決定処理では、図３のような木構造に基づいて分割位置が決定される。こうして原文のフレーズ境界が決定され、次に、図２に示すような原文単語と訳文単語との対応情報に基づいて、各フレーズに属する訳文単語が求められる。そして、各フレーズにおいて、そこに属する訳文単語を、全文生成の場合の生成順として定められた順番の昇順に詰めて、各フレーズの訳のフレーズが得られる。

最後に、図４のように、原語 - 訳語対応情報をフレーズ単位で付け直す。この結果、この翻訳エンジンを利

英単語番号	英単語	対応する訳文単語番号
1	In	2
2	addition	2
3	to	2
4	sequential	1
5	files,	1
6	the	7
7	second	8
8	file	9
9	structure	10
10	used	6
11	by	5
12	SHARP-BASIC	4
13	is	14
14	random	13
15	access.	13

訳文単語番号	訳文単語	対応する英単語番号
1	逐次ファイル	4
2	に加えて	1
3	、	なし
4	シャープ-BASIC	12
5	によって	11
6	使われる	10
7	(訳語なし)	6
8	第2の	7
9	ファイル	8
10	構造	9
11	は	なし
12	、	なし
13	ランダムアクセス	14
14	である	13
15	。	なし

図 2: 原文単語 - 訳文単語対応情報

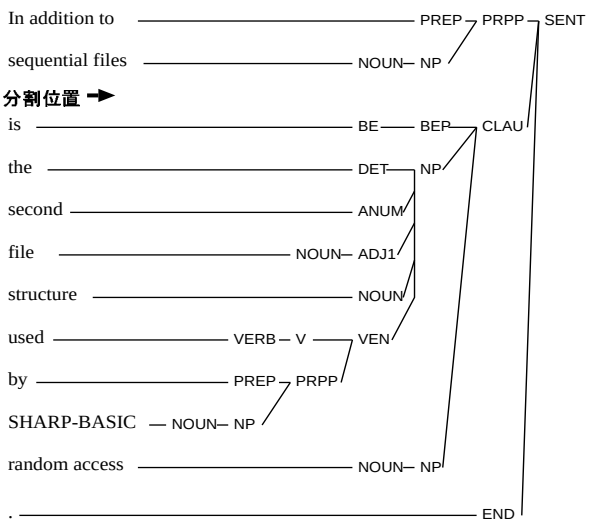

図 3: 木構造と分割位置

英単語番号	英単語	対応する訳文フレーズ番号
1	In	1
2	addition	1
3	to	1
4	sequential	1
5	files,	1
6	the	2
7	second	2
8	file	2
9	structure	2
10	used	2
11	by	2
12	SHARP-BASIC	2
13	is	2
14	random	2
15	access.	2

訳文フレーズ番号	訳文フレーズ	対応する英単語番号
1	逐次ファイルに加えて	1
2	シャープ-BASICによって使われる第2のファイル構造は、ランダムアクセスである	6

図 4: 原文単語 - 訳文フレーズ対応情報

In addition to sequential files, the second file structure used
逐次ファイルに加えて シャープ-BASIC によって使われ

by SHARP-BASIC is random access.
る第2のファイル構造は、ランダムアクセスである

図 5: フレーズ訳の出力イメージ

用して作成された翻訳システムでは、表示の際に図５のように各フレーズの頭を原文と訳文とで揃えて出すことが可能になる。

3　フレーズ訳の生成

本節では、上記の「日本語生成」処理中に実行される、フレーズ訳の生成の基本的な手順を説明する。

3.1　原文の分割

まず、原文のフレーズへの分割位置が決定される。図２のような木構造の各々の節点を所定の順序で探索し、各々の節点で、図６に示したような分割規則が適用可能か否かを調べ、可能なら分割位置設定の措置を行う。

この段階で、適用する規則を適宜選択することによって、分割箇所を少なめにして長めのフレーズとしてまとめるか、逆に分割箇所を多めにして短めのフレーズにするか、の調整をすることが可能である。

規則番号	分割規則
1	SENT の直下が (END を除いて) 複数のノードになっていたら、分割する。
2	CLAUSE 系のノードは分離する。

図 6: 分割位置決定規則の例

フレーズ番号	原文単語範囲
1	1 — 5
2	6 — 15

図 7: 各フレーズの範囲

3.2 訳文単語の各フレーズへの割り当て

図２の木構造の図に示したように分割位置が決定された結果、原文における各フレーズの範囲が決まる。これらを原文の単語位置で示したのが図７である。こうして原文の各々の単語がどのフレーズに属するかが求まるので、各々の原文単語に対して図２のような原文単語 - 訳文単語対応情報を参照した結果、ここに対応関係が示されている訳文単語については、それぞれどのフレーズに属するかが求まる。

3.3 対応する語が原文にない訳語の割り当て

次に、図２のような原文単語 - 訳文単語対応情報に対応関係が示されていない訳文単語がどのフレーズに属するかを求める。

このような訳文単語は、多くの場合は助詞などの付属語であると考えられる。そこで、付属語であるなら、それと同一の文節に属する自立語と同一の訳文フレーズに属すべきであるという考え方を原則とした。ここでは、原文単語 - 訳文単語対応情報に対応関係が示されていない一つの訳文単語について、その訳文単語の（文として生成した場合の順序において）直前にあって、原文単語 - 訳文単語対応情報に対応関係が示されているような訳文単語（自立語である可能性が高いと考えられる）と同一の訳文フレーズに属するとみなす。たとえば、図２において 11 番目の訳文単語「は」は、10 番目の訳文単語「構造」と同一のフレーズに属するとする。

図２のような関係の文に関して訳文単語を各フレーズに振り分けて、フレーズごとに訳文単語の順序をソートした結果が図８である。この結果から、各フレーズにおいて訳文単語を並べて、フレーズ訳を生成した結果が図４のようになる。

フレーズ番号	訳文単語番号
1	1, 2
2	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

図 8: 各フレーズに属する訳文単語番号

4 訳の自然さの向上のための調整

著者らは、前節までの方針で構築したフレーズ訳生成機能を英日機械翻訳システムにいちど実装したが、さらに読みやすさを増すために、その後、以下のような改善を行った。

4.1 助動詞由来の様相の、各節の末尾の訳出への反映

たとえば、

He can sing a song and write a song.

という英文を前節までの方針で訳すと、原文の分割結果が

1. He can sing a song
2. and
3. write a song

のようになり、フレーズ訳を生成した結果が

1. 彼は、歌を歌い
2. そして
3. 歌を書くことができる

となる。この例では、原文と翻訳結果とをフレーズ単位で比較すると、1. のフレーズにおいては原文に助動詞 can が存在するが、このことが 1. の訳文フレーズには反映されていない。助動詞 can の効果は 3. の訳文フレーズのみに「〜ことができる」という形で現れている。この原因は、原文を１文単位で通常翻訳をした結果が

彼は、歌を歌い、そして、歌を書くことができる。

というようになり、助動詞の効果が文末のみに現れていることと、原文のフレーズ分割位置を忠実に訳文に反映させたためである。

しかし、1. のフレーズのみで原文と訳文とを比較してみると、原文には助動詞が存在するのに訳文にはその効果が現れていないのは不自然である。原文において、原文の助動詞 can は、主語を共有して並列になっている原文の２つの節の両方に対して効いている。

そこで、このような場合、フレーズ翻訳ではそれぞれの節に対して助動詞の効果が及んでいることを明示的に表し、

1. 彼は、歌を歌うことができ
2. そして
3. 歌を書くことができる

のように節末の訳の形を変えるようにした。

4.2 格の後置語の訳出位置の調整および補助記号の生成

たとえば、

I said that he ate it.

という英文を前節までの方針で訳すと、原文の分割結果が

1. I said
2. that he ate it.

のようになり、フレーズ訳を生成した結果が

1. 私は、言った
2. 彼がそれを食べると

となる。通常の文生成をすると「私は、彼がそれを食べると言った。」となるが、フレーズ訳ではこの全文から「彼がそれを食べると」のフレーズが中抜けしたような形となっている。

この例では、原文単語と訳文単語との対応関係は図９のようになり、格助詞「と」は原文単語との対応関係がないとされるため、その直前の訳文単語「食べる」と同一の訳文フレーズに属するようにしている。しかし、このようにフレーズが中抜けし、その抜けたフレーズの末尾が格助詞であるような場合は、1. のフレーズを

私は、～と言った

のように、抜けた元の箇所にも格助詞を補い、抜けたことがわかるように「～」を生成するようにした。

英単語番号	英単語	対応する訳文単語番号
1	I	1
2	said	10
3	that	なし
4	he	4
5	ate	8
6	it.	6

対応関係

訳文単語番号	訳文単語	対応する英単語番号
1	私	1
2	は	なし
3	、	なし
4	彼	4
5	が	なし
6	それ	6
7	を	なし
8	食べる	5
9	と	なし
10	言った	2
11	。	なし

図 9: 原文単語 - 訳文単語対応情報

4.3 関係節への先行詞の訳出

関係節を含む文の場合、前節までの方針でフレーズ訳を得ると、原文は、

It contains some impurities which we remove by smelting.

原文のフレーズへの分割結果は、

1. It contains some impurities
2. which we remove by smelting.

フレーズの翻訳結果は、

1. それは、いくらかの不純物を含む
2. 我々が溶解によって除去する

のように、関係節のフレーズの訳は関係詞にあたっている格が欠けた訳となる。そこで、主節にある先行詞を、関係節にも補い、関係節の訳フレーズが

その不純物を我々が溶解によって除去する

となるようにした。

このように関係節の欠落した格を補うためには、格の後置語（格助詞など)、先行詞、補助修飾語を推定する必要がある。

- 格の後置語は、フレーズ生成の時点で得られている関係節の格構造から推定する。推定不可能な場合は「は」とする。
- 先行詞は、木構造において関係詞の親の節点より上位にあって、それと同じ種類である節点を頂点とする部分木のヘッドである。この例では、図１０において、which の親の節点が NP(名詞句) で、これより上位にある NP を頂点とする部分木とい

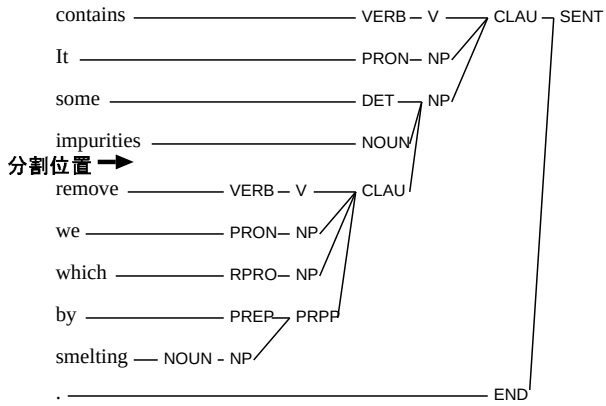

図 10: 関係節を含む文の木構造

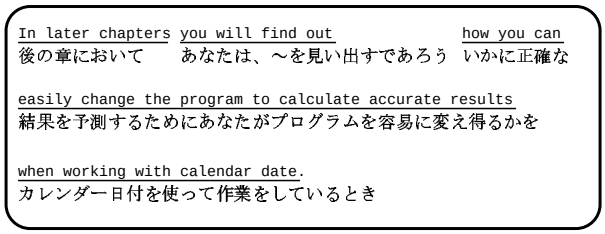

図 11: フレーズ訳の出力イメージ

うことで、“ some impurities which we remove by smelting ” をカバーする部分木がそれにあたり、これのヘッドが impurities と定められている。

- 補助修飾語は、現時点では「その」とする。

5 おわりに

上記のように、英日機械翻訳エンジンに付加する形でのフレーズ訳生成機能の開発と改善を行った。この結果、係り受けの複雑な文でも、図 11 に示すように、読みやすい形で翻訳結果を出力することができる。この成果は 1996 年 3 月にシャープから発売予定のノートパソコンにバンドルされる英日機械翻訳システムに搭載されている。

今後は、漸進的な解析手法なども含めて、一文単位にとらわれない翻訳手法の研究はますます発展していくと考えられる。フレーズ訳生成については、更に語句・記号を補うことによる読みやすさの向上をメインとして今後も改善していきたいと考えている。