

日本語・英語文パーザを用いた日英文型パターン対の半自動構築

小林久敏

宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

日英文型パターン対を用いた単文翻訳システムでは、入力文から抽出された単文を主体表現と客体表現に分離し、客体表現から英語文の骨格を作成し、その結果と主体表現の情報を組み合わせることによって日本語単文に対応する英語文を生成する。そのため、技術文を対象とする高品質な日英機械翻訳システムを実現するには、各分野特有の表現に関する日英文型パターン対が必要となる。

本稿では、このような専門分野の日英文型パターン対を効率的に構築するものとして、日英対訳コーパスを日本語文や英語文パーザで構造化し、日英間で対応する単文対を抽出すると共に、単文内で句や単語を対応づける方法を提案し、その有効性について論じる。

2 日英文型パターン対を用いた翻訳

文型パターン辞書には約 15000 件の一般語を対象とした日英対訳パターン [1] がすでに人手により作成されている。表 1 にその例を示す。

日英文型パターン対を用いた翻訳システムでは、まず入力文の述語をキーにして文型パターン対を検索し、入力文の格パターンと一致した文型パターン対の中から入力文の名詞意味属性と最も近いものを選択し、英語文の骨格を生成する。

表 1: 日本語語彙大系の日英文型パターン対

表記	流す
日本語パターン	N1 が N2 を 流す
英語パターン	N1 carry away N2
名詞意味属性	N1(*) N2(施設 人工物)

3 日英単文対作成

技術文は専門分野に固有の専門用語及び用法があり、一般語を対象とした文型パターン対では望ましい翻訳結果は期待できない。そのため、技術文の対訳コーパスから日英文型パターン対を生成することで技術文の翻訳に適應できる。

日英文型パターン対を構築するためには日英の単文対が必要となる。まず文単位に分割した日英対訳コーパスを既存のパーザで構造化し、日本語節、英語節を抽出する。そして、その抽出された日英節の用言から対応する節を検索し、日英単文対を生成する。

なお、技術文を対象とするため、約 15000 文の技術文の日英対訳コーパスをデータとして使用した。

3.1 日本語節抽出処理

日本語節抽出には形態素解析結果を用いる。形態素解析には日本語形態素解析システム Jumper[2] を使用する。形態素解析結果を動詞句文節、形容詞文節、名詞文節 (文末、形容動詞語幹+だ)

の箇所でも自動分割する。また、述部文節頭部の自立語までを節として取得する。

表 2: 名詞、形容詞節の自動分割箇所

分割箇所		例
品詞	活用形	
形容詞	未然形	精度が良く/ない
	連用形	精度が良く/、
	終止形	精度が良い/。
	假定形	精度が良けれ/ば
名詞	文末	不正に頑健/だ。
	形容動詞語幹+で	不正に頑健/で、

原文

入力文を解析するプログラムは精度が良く、不正に頑健だ。

形態素解析結果

入力文 (\$N)/を (pP12)|解析 (NVS)/する (vSA7)|プログラム (NNG)/は (pK11)|精度 (NNG)/が (pP11)|良 (AUA0)/く (jUA3)/、 (SC1)|不正 (NAD3)/に (pP13)|頑健 (NAD3)/だ (xa16)/。(SP1)

節抽出結果

- ・ 入力文を解析する
- ・ プログラムは精度が良い
- ・ 不正に頑健だ

3.2 英語節抽出処理

英語節抽出には構文解析結果を用いる。構文解析には英語構文解析システム Charniak[3]を使用する。解析結果から木構造が得られるので、節相

当 (埋込み文も含む) の部分木を抽出できる。

原文

The element is manufactured
by sandwiching an compound between
electrodes.

構文解析結果

```
(S1 (S (NP (DT The) (NN element))
(VP (AUX is) (VP (VBN manufactured)
(PP (IN by) (S (VP (VBG sandwiching)
(NP (NP (DT an) (NN compound))
(PP (IN between)
(NP (CD electrodes)))))))))))))
```

節抽出結果

- ・ the element is manufactured.
- ・ sandwiching an compound between electrodes

3.3 日英節対応処理

日本語節、英語節それぞれの用言に対して日英対照辞書を検索し、その対応から日英単文対を生成する。ここで、日本語、英語の屈折変化を吸収する。また日英の用言に関して交差検定を行い、網羅性を確認する。

一般語の日英対照辞書では技術文特有の動詞表現が収録されていないので、専門的な用言を含む単文対が抽出されない。そこで、英語の埋込み文、日本語の名詞修飾文の関係から節対応を推測する。その結果から日英動詞対を抽出し、頻度等の情報を付与することによって、技術文特有の動詞辞書を自動生成する。これにより、一般語の辞書では得ることができない単文対の抽出が可能となる。

表 3: 自動抽出した日英動詞対の例

積層する	stack
逸脱する	depart
印加する	apply
駆動する	drive
照射する	irradiate
劣化する	deteriorate
析出する	filtrate
結晶化する	crystallize

3.4 日英単文対応処理結果

約 15000 文の技術文から単文対応処理を行った。日本語節 45000、英語節 35000 の中から 15000 対の単文対が抽出された。

また、抽出された単文対における述語の再現率を調べるため 15000 対の単文対を 500 対程度に分割し、交差検定を行うと、再現率が 94%という高い網羅性を示した。

・単文対応が失敗する場合

日英間で単文対を作成する際、無生物主語構文 [4] のように日本語と英語が非線形対応なものは、自動処理では対応をとることができない。無生物主語構文とは述部に動作を表す動詞を使っているが、主語が意志をもたない無生物である構文のことをいい、因果関係を記述する文に使われるのが一般的で、技術文には頻繁に用いられる。

下記の例 [5] では、日本語側の用言が「送信する」であるのに対し、英語側の用言は「allow(許す)」となっており、用言の対応がとれず単文対が抽出されない。

・失敗例

このアンテナにより、映像信号と音声信号の両方を送信できる。

This antenna allows the transmission of the video and sound signals.

4 日英文型パターン対作成

文型パターン対を構築するためには、その動詞の格パターンを決定しなければならない。抽出した単文対から格パターンを生成するため、名詞辞書を用いて単文対内の句や単語の対応づけを行う。しかし、専門用語は各分野特有の用語であり、また名詞句や複合名詞などは表現が多様であるので一般語の辞書で対応づけを行うことは難しい。そこでまず、一般語の辞書から技術文に対する複合名詞辞書を自動生成する。その辞書を用いて日英単文対内の名詞、名詞句を対応づけ、格パターンを決定し、文型パターン対を構築する。

4.1 日英名詞対応処理

日本語文は、日本語形態素解析システム Jamper で形態素解析 [1]、英語文は英語構文解析システム Charniak[3] で構文解析を行い、文節内の主名詞 (日本語文)、名詞句 (英語文) を抽出する。日本語文は名詞、複合名詞、英語文は名詞句が抽出され、英語名詞句の各単語の訳語から日本語の名詞、複合名詞との対応をとり、日英名詞対を抽出する。さらに、頻度等の情報を付与し、技術文に対する名詞辞書を作成する。

表 4: 対応名詞対例

薄膜トランジスタ	thin film transistor
半導体装置	semiconductor device
発光素子	light emitting element
ゲート電極	gate electrode
インジウム錫酸化物	indium tin oxide
無線 I C タグ	wireless IC tag
高密度プラズマ処理	high density plasma process
結晶性半導体膜	crystalline semiconductor film
レーザ照射装置	laser irradiation apparatus
有機化合物	organic compound
ドレイン電極層	drain electrode layer

4.2 日英名詞対応処理結果

15000 文の日英対訳コーパスから約 30000 対の日英名詞対応が得られた。

しかし、下記の例のように品詞の変換が生じている文 [5] だと、日本語側の「回転効率」という名詞に対応する英語側の名詞 (句) が存在せず、名詞対を抽出することができない。

・名詞対応失敗例

定期的な保守により、モーターの回転効率が高まる。

Periodical maintenance enables the motor to rotate more efficiently.

4.3 パターン生成

抽出された単文対からそれぞれの格要素を汎化し、格パターンを生成する。汎化された日英の名詞 (句) を対応づけ、文型パターン対を構築する。

技術文において、日本語は人間を主語とし、省略するのが一般的であるが、英語は人間ではなく物や事柄を主語として記述することが一般的である。よって、文型パターン対を構築する際、態情報も重要な情報となる。[1] の日英文型パターン対では日本語パターンは能動態で書かれていた。しかし、技術文では日本語文で受動態が自然な場合には、日本語側の文型パターンは受動態で記述し、態情報を記述することにした。

5 まとめ

今回の方法では、約 9000 対の文型パターン対が自動抽出された。計算機における自動抽出なので必ずしも適した文型パターン対とは限らない。実際の翻訳システムに反映させるためにはこの文型パターン対を目視確認し、洗練させる必要がある。

また、半自動生成した日英文型パターン対は日本語語彙大系における日英文型パターン対と異な

り、名詞意味属性の情報が付与されていない。半自動生成した日英文型パターン対にどのような名詞意味属性を付与するか今後検討が必要である。

表 5: 日英文型パターン対の例

日本語動詞	結晶化する
英語動詞	crystallize
日本語格パターン	N1 が N2 で 結晶化する
英語格パターン	N1 be crystallized with N2
日本語態	能動
英語態	受動
日本語名詞	N1(*), N2(非晶質珪素膜)
英語名詞 (句)	N1(*), N2(amorphous silicon film)

日本語動詞	挟む
英語動詞	sandwich
日本語格パターン	N1 が N2 に N3 を 挟む
英語格パターン	N1 sandwich N3 between N2
日本語態	能動
英語態	能動
日本語名詞	N1(*), N2(一対の電極間) N3(有機化合物)
英語名詞句	N1(*),N2(a pair of electrodes) N3(organic compound)

参考文献

- [1] 池原悟、宮崎正弘、白井諭、横尾明男、中岩浩巳、小倉健太郎、大山芳史、林良彦：日本語語彙大系、岩波書店、1997
- [2] 宮崎正弘、川辺諭、武本裕：構造化チャート法に基づく日本語形態素解析器 jamper 言語処理学会第 13 回年次発表論文集、2007
- [3] E. Charniak: A Maximum-Entropy-Inspired Parser. Technical Report CS-99-12, 1999
- [4] 富井篤：技術翻訳のテクニック、丸善株式会社、1993
- [5] 佐藤洋一：技術英文大全 (第 3 巻) 高度な英文表現、オーム社、2005