E_029

関係節の英日翻訳法に関する統計的特徴抽出の試み
Statistical Acquisition of the Knowledge for English-Japanese Translation of Relative Clauses

九津見 毅* 吉見 毅彦† 小谷 克則‡ 飯田 将人* 佐田 いち子* 井佐原 均‡
Takeshi KUTSUMI Takehiko YOSHIMI Katsunori KOTANI Masato IIDA Ichiko SATA Hitoshi ISAHARA

1 はじめに

英語-日本語間のように構文構造の大きく異なる言語間の翻訳においても、原文の語順をなるべく保ったまま目的言語の文に翻訳することが望ましいという傾向が強まっている。これに対し、機械翻訳システム開発者は、たとえば文献（佐田 九津見 日野 関谷 1998）に示すように、複文の訳し方を工夫するといった対応をしてきた。特に、訳し方の順序が「わかりやすさ」に大きくかかわってくる関係節に関しては、従来の英文和訳的な「関係節を前置修飾する」の他に、たとえば文献（九津見 奥西 佐田 1996）などに示すように、

- 関係節を別文として主節の後に生成する
- 関係節を括弧に入れて先行詞の直後生成する

といった取り組みがなされてきた。たとえば、

英語文 It contains some impurities which we remove by smelting.

従来の日本語訳 それは、我々が溶解によって除去するいくらかの不純物を含む。

改良日本語訳 それは、いくらかの不純物を含む。その不純物を我々が溶解によって除去する。

しかし、現状では次のような問題点がある。

(1) 人間の優秀な翻訳者による翻訳と比べると、表現方法のバリエーションが圧倒的に少なく、節のつなぎの箇所の処理方法などに非常に稚拙な印象を与える。
(2) どのような場合に、どのような訳し方をすべきかの判断基準が十分に解明されていない。

我々の研究では、科学技術論文の英日対訳コーパスを使用し、関係節について人間翻訳者の訳し方をパターン化し機械翻訳に採り入れることを目指している。本稿では、関係節を含む英語文の特徴を素性とし、関係節を含む日本語文の特徴をクラスとして、素性の組み合わせからクラスを予測することの可能性を探る。

2 研究方法の概要

本稿では、関係節を含む文の英日対訳パターンを獲得するための過程として、関係節を含む英日対訳文の英語文の特徴（素性）と日本語文の特徴（クラス）との関連を分析することを目指す。研究の焦点を絞るため、本稿では、関係代名詞 "which" によって導かれる関係節のみを対象とする。

手順は以下の通りである。

(1) 文アライメントされた英日対訳コーパスから、英語文に "which" を含む対訳文の組を抽出する。このうち、"which" が関係代名詞でないものを対象外とする。
(2) 対訳文のうち英語文の特徴（素性）を判定する。本稿においては4種類の素性を採用し、判定をプログラムで自動的に行った。
(3) 対訳文のうち日本語文の特徴（クラス）を判定する。本稿においては8種類のクラスを採用し、判定を人手で行った。
(4) 各クラスごとに、素性の組み合わせとの多変量解析を行って、統計モデルを求める。この統計モデルによるクラス分け結果を、交差検定により評価する。

* シャープ(株), Sharp Corp.
† 龍谷大学/情報通信研究機構, Ryukoku Univ. / NICT
‡ 情報通信研究機構, NICT

3 関係節を含む英日対訳文の条件とクラス

本節では、実験に採用した英語文の特徴（素性）と日本語文の特徴（クラス）について述べる。

3.1 条件部——英語文の特徴

関係節を含む英語文の特徴を、次の4種類の観点から分類する。

(1) 関係詞のカンマや前置詞の有無
(2) 関係詞の関係節中における格
(3) 関係節の文中における位置
(4) 関係節の長さ(単語数)

それぞれの素性の分類法を、以下に示す。

(1) 関係詞のカンマや前置詞の有無
[1-1]単純関係詞
[1-2]前置詞付き関係詞
[1-3]カンマ付き関係詞
[1-4]カンマと前置詞付き関係詞
[1-5]カンマ及び数量を表す語と前置詞付き関係詞
[1-6]上記以外

(2) 関係詞の関係節中における格
[2-1] 関係詞が関係節の主語
[2-2] 関係詞が関係節の主語以外の動詞格
[2-3] 関係詞が関係節の前置詞目的語
[2-4] 関係詞が関係節の受身主語
[2-5] 上記以外

(3) 関係節の文中における位置
[3-1] 関係節が文末
[3-2] 関係節が文の途中かつ節の途中
[3-3] 関係節が文の途中だが節末
[3-4] 一文全体が名詞句で、関係節が文末

(4) 関係節の長さ(単語数)
[4-1] 4単語以下

[4-2] 5～9単語
[4-3] 10～14単語
[4-4] 15～19単語
[4-5] 20～24単語
[4-6] 25～29単語
[4-7] 30単語以上

3.2 クラス部——日本語文の特徴

関係節を含む日本語文の特徴を、次の8種類の観点から分類する。

(A1) 関係節の生成位置（上位節との関係で）
(A2) 関係節と上位節との構文構造[1]
(B) 関係節の、関係詞格要素の扱い
(C) 関係節の末尾の扱い
(D) 上位節の末尾の扱い
(E) 上位節の、末尾以外の箇所
(F) 上位節の全体的な変更
(G) 先行詞の変更

それぞれのクラスの分類法（群）を、以下に示す。

(A1) 関係節の生成位置（上位節との関係で）
[A1-1] 関係節が上位節の内部にある
[A1-2] 関係節が上位節と分離している
[A1-3] 日本語文に関係節相当部分がない
[A1-4] 先行詞が同格並列であり、その2つ（あるいはそれ以上）の先行詞の訳され方が異なる
[A1-o] その他

(A2) 関係節と上位節との構文構造
[A2-1] 関係節が前置修飾
[A2-2] 関係節が上位節の後置の節（別文生成の場合を含む）
[A2-3] 関係節が、上位節の主語を格要素とする節になっている
[A2-4] 関係節が、名詞句化した上位節を格要素とする節になっている
[A2-5] 関係節にさらに関係節や従属節が付随していて、上位の関係節は主節内に前置修飾だが、下位の関係節や従属節は主節と分離して後置
[A2-o] その他

(B) 関係節の、関係詞格要素の扱い
[B-1] 何もしない
[B-2] 関係節における関係詞格要素を何らかの形で補う
[B-3] 英語文の関係代名詞が前置詞付きであり、それが関係節中に訳出されている
[B-4] 先行詞が関係節の内部にある。
[B-o] その他

(C) 関係節の末尾の扱い
[C-1] 終止形にする
[C-2] 連体修飾にする
[C-3] 連用中止にする
[C-4] 形態素を補う
[C-5] 変更する
[C-o] その他

(D) 上位節の末尾の扱い
[D-1] 文末にする
[D-2] 連用中止にする
[D-3] 形態素を補う。
[D-4] 文末ではないが、節が終止形になっている
[D-o] その他

(E) 上位節の、末尾以外の箇所
[E-1] 特に何もしない
[E-2] 末尾以外の箇所に形態素を補う
[E-3] 上位節全体を名詞句化
[E-o] その他

(F) 上位節の全体的な変更
[F-1] 特に何もしない
[F-2] 目的語を中心とする名詞句（主語・動詞は目的語にかかる関係節）にする
[F-o] その他

(G) 先行詞の変更
[G-1] 特に何もしない
[G-2] 変更する
[G-o] その他

4 実験

4.1 実験の概要

実験には、ネイチャー・ジャパン株式会社の許諾を受けて同社より提供を受けた、英語科学技術論文のアブストラクトとその日本語訳文との対訳コーパスを用いた。英日文の文単位アライメントはシャープ（株）において行った。

本研究で対象とした文対応付け対訳コーパスの文数は7,115組であるが、そこから、文頭以外に “which” を含む文を抽出した結果、550組が抽出された。このうち、英日文アライメントの誤りや “which” が関係代名詞でないなどの理由で評価対象外とした文を除いた、評価対象文数は533組で、対象とするwhich関係節の個数は553個であった（一文中にwhich関係節が複数個含まれる場合もあるので）。

[1]当初、(A1)と(A2)は分けずに「日本語文における関係節と上位節との構文構造」としていたが、「構文構造と、単純な位置関係とは、分けて扱うべき」という意見を考慮し、現状のように設定した。

英語文の関係節の特徴抽出においては、シャープ英日翻訳システムの構文解析系を改造したプログラムにより、関係節とその特徴を検出した。[2]

日本語文の関係節の特徴抽出においては、評価者により人手で判断した。

4.2 個別の集計結果

英語文の各素性及び日本語文の各クラスを個別に集計した結果を、表1～12に示す。

表1　素性1の集計結果

1	2	3	4	5	6
93	82	299	21	20	38

表2　素性2の集計結果

1	2	3	4	5
301	21	153	71	7

表3　素性3の集計結果

1	2	3	4
378	4	165	6

表4　素性4の集計結果

1	2	3	4	5	6	7
57	248	124	68	24	15	17

表5　クラスA1の集計結果

1	2	3	4	o
263	280	3	3	4

表6　クラスA2の集計結果

1	2	3	4	5	o
194	252	54	11	1	41

表7　クラスBの集計結果

1	2	3	4	o
365	148	22	6	12

表8　クラスCの集計結果

1	2	3	4	5	o
253	174	43	61	3	19

表9　クラスDの集計結果

1	2	3	4	o
322	146	66	5	14

表10　クラスEの集計結果

1	2	3	o
508	29	6	10

表11　クラスFの集計結果

1	2	o
532	8	13

表12　クラスGの集計結果

1	2	o
524	18	11

5 統計的検定と考察

本稿では、英語文の4種の素性の組み合わせと、日本語文の個々のクラスとの関係を統計モデル化する手法として、林の数量化理論第II類を採用する[3]。こうして得られた（個々のクラスごとの）判別関数は、英語文の素性（英語関係節の特徴）から、その特徴に合った日本語文のクラスの値（関係節の訳し方）を求めるものになる。

評価方法として、5分割交差検定法を用いた。訓練データにより判別関数を求め、実験データの英語文の素性からその判別関数で求めた日本語文のクラスの値が観測値と合致しているかを調べた。

合致しているかの判断は以下のように行った。訓練データにより判別関数を求める際に、訓練データにおける群（クラスの値）ごとの判別得点の平均値を求め、判別得点平均値が隣接している2群間において、その隣接する2群それぞれの判別得点平均値の中間値を、隣接する2群のどちらに属するかの境界値とした。実験データを判別関数に適用した判別得点が、上記のようにして分割された群のどれに属するかを、判別関数により得られたクラスの値とし、それが実際に観測されたクラスの値と合致しているかを調べた。

その結果から各クラスの各群について正解数・漏れ数・誤り数を集計し、そこから群ごとの適合率・再現率・F値を求めた。その結果を表13に示す。

表13から次のことが言える。

判別結果が良好である場合もある。たとえば、クラスB(関係節の関係詞格要素の扱い)の判定結果が3（英語文の関係代名詞が前置詞付きであり、それが関係節中に訳出されている）の場合は、適合率も再現率も高く、F値も高い値を示している。

また、クラスE(上位節の末尾以外の箇所)の判定結果が1（何もしない）の場合や、クラスF(上位節の全体的な変更)の判定結果が1（何もしない）の場合、クラスG(先行詞変更)の判定結果が1（何もしない）の場合などは、適合率が高い値を示している。だが、表10～表12からわかるように、これらのクラスのとる値はもともと1である確率が圧倒的に高い。とりあえず1としておけば間違いは少なく、もともと判別が容易な状況であると言える。

[2]英語文の関係節の検出には、誤検出や検出漏れがあり得る。この対策として、人手による日本語文の関係節の検出結果と比較して、文中に存在する関係節の個数が日英で異なる場合は、明らかに誤検出が起こっているとみなして、英語文の特徴抽出も人手で判定した。

[3]日本語文の特徴クラスについては、クラス間の依存関係がみられる可能性があることから、多次元のクラスを個別に検討するのではなく組み合わせクラスとして1次元のクラスに変換し、それと英語文の素性の組み合わせとで統計モデルを求めることも試みた。しかし、その場合、クラスの値が百数十通りとなり、数量化理論第II類などの手法では有効な統計モデルを求めることができなかった。

上記以外の場合では、総じて、判別成績は良いとは言えない。

表13　数量化理論第II類による統計モデルを交差検定した実験結果

クラス	群	正解	漏れ	誤り	適合率	再現率	F値
A1	1	20	243	22	0.4762	0.0760	0.1311
	2	118	162	79	0.5990	0.4214	0.4948
	3	0	3	171	0.0000	0.0000	—
	4	0	3	33	0.0000	0.0000	—
	o	1	3	109	0.0091	0.2500	0.0175
	計	139	414	414	0.2514	0.2514	0.2514
A2	1	73	121	95	0.4345	0.3763	0.4033
	2	74	178	66	0.5286	0.2937	0.3776
	3	6	48	51	0.1053	0.1111	0.1081
	4	0	8	117	0.0000	0.0000	—
	5	0	1	4	0.0000	0.0000	—
	o	3	38	61	0.0469	0.0732	0.0571
	計	156	394	394	0.2836	0.2836	0.2836
B	1	39	326	18	0.6842	0.1068	0.1848
	2	60	88	140	0.3000	0.4054	0.3448
	3	16	6	5	0.7619	0.7273	0.7442
	4	0	6	100	0.0000	0.0000	—
	o	4	8	171	0.0229	0.3333	0.0428
	計	119	434	434	0.2152	0.2152	0.2152
C	1	86	167	78	0.5244	0.3399	0.4125
	2	25	149	52	0.3247	0.1437	0.1992
	3	11	32	110	0.0909	0.2558	0.1341
	4	12	49	83	0.1263	0.1967	0.1538
	5	0	3	71	0.0000	0.0000	—
	o	1	18	24	0.0400	0.0526	0.0455
	計	135	418	418	0.2441	0.2441	0.2441
D	1	30	292	19	0.6122	0.0932	0.1617
	2	50	96	105	0.3226	0.3425	0.3322
	3	13	53	105	0.1102	0.1970	0.1413
	4	0	5	90	0.0000	0.0000	—
	o	1	13	140	0.0071	0.0714	0.0129
	計	94	459	459	0.1700	0.1700	0.1700
E	1	199	309	19	0.9128	0.3917	0.5482
	2	4	25	95	0.0404	0.1379	0.0625
	3	0	6	55	0.0000	0.0000	—
	o	2	8	179	0.0110	0.2000	0.0209
	計	205	348	348	0.3707	0.3707	0.3707
F	1	197	335	7	0.9657	0.3703	0.5353
	2	1	7	156	0.0064	0.1250	0.0121
	o	4	9	188	0.0208	0.3077	0.0390
	計	202	351	351	0.3653	0.3653	0.3653
G	1	288	236	15	0.9505	0.5496	0.6965
	2	6	12	144	0.0400	0.3333	0.0714
	o	5	6	95	0.0500	0.4545	0.0901
	計	299	254	254	0.5407	0.5407	0.5407

6 おわりに

本稿では、英語の関係節の自然な日本語への翻訳方法を獲得するための手段として、英日対訳コーパスを用いて、関係節を含む英語文の特徴（素性）と日本語文の特徴（クラス）を多次元的に抽出し、素性の組み合わせからクラスの値を統計的に判別することの可能性についての検討を行った。

科学技術論文のアブストラクトを対象として、which関係節を含む対訳文533組（which関係節の個数553個）を調査し、数量化理論第II類で統計モデル化を行い、交差検定で評価した結果、特定のクラス（例として、関係節の関係詞格要素の扱い）については良好に判別できた。

さらに予測性能を向上させるために、以下のような取り組みを目指していく。

・利用する英語文の特徴の見直し
・他の解析手法の採用検討

7 謝辞

本研究に用いた英日対訳コーパスは、シャープ株式会社がネイチャー・ジャパン株式会社の許諾を受けてネイチャー・ジャパン株式会社より提供を受けた、英語科学技術論文のアブストラクト及び、その日本語訳文を、シャープ株式会社において文単位アライメントを行って英日対訳文コーパスとしたものである。

同コーパスの研究目的への利用を許諾いただいたネイチャー・ジャパン株式会社に深謝する。

参考文献

九津見毅，奥西稔幸，佐田いち子．"英日機械翻訳における速読支援のための日本語生成"．言語処理学会第2回年次大会発表論文集 pp.221-224．(1996)

佐田いち子，九津見毅，日野ちなみ，関谷正明．"英文語順に準拠した日本語生成"．言語処理学会第4回年次大会発表論文集 pp.670-672．(1998)

Timothy Baldwin, 田中穂積, 徳永健伸. "日本語の関係節における主辞の省略の解析"．情報処理学会研究報告 NL117-1 (1997).Alderson, J. C. 2000. *Assessing Reading*. Cambridge University Press, Cambridge.