ソフトウェアマニュアルを用いた係り受けコーパスの試作

佐藤 美沙　　柳井 孝介
（株）日立製作所 研究開発グループ
{misa.sato.mw, kohsuke.yanai.cs}@hitachi.com

1 はじめに

係り受けの解析結果は情報抽出 [3] などの多くの応用領域で活用されている。自然言語処理分野では新聞記事などを元にした係り受けコーパス [7] が整備されており、それを用いた教師あり学習により高精度な係り受け解析が行えることが知られている [4, 6]。

既存の係り受けコーパスは新聞記事や一般書籍を用いたものが中心であり [2, 5, 7, 8]、ソフトウェアの製品マニュアルのような、個別具体的な技術情報について書かれた文書では作られていない。こうした技術文書は新聞記事や一般書籍とは異なる構文的特徴を持つと考えられるため、製品マニュアルを用いた係り受けコーパスが用意されれば、技術文書に対して、より精度の高い係り受け解析を行える可能性がある。

製品マニュアルの内容を正確に理解するには、製品についての専門知識が必要となる。また、係り受けの付与は専門性の高い作業であるため、言語についての専門的な知識を持って行う必要がある。しかしながら、係り受け付与作業者として、製品知識を持つ技術者であり、かつ言語の専門家である者を確保することは非常に困難である。

そこで本稿では、製品マニュアルを用いた係り受けコーパスについて、製品知識がなく、係り受け付与作業経験も持たない作業者による試作の結果を報告する。

2 係り受けコーパスの構築

2.1 係り受けの判断基準

文節および係り受けは、京都大学テキストコーパス [7] の判断基準に従う。代表的な係り元と係り先の関係は、格要素から述語、修飾語から被修飾語、従属節の述語から主節の述語、並列、同格である。京都大学テキストコーパスでは、並列と同格の係り受けに対しては他の係り受けと区別してラベルを付与しているが、本稿では区別せずに扱う。

2.2 係り受けコーパス作成手順

係り受けコーパスの作成は、(1) 文の収集、(2) 係り受けテンプレートの作成、(3) 手作業による修正、(4) 形態素情報との結合、の手順で行う。図 1 に作成手順と例を示す。

(1) 文の収集　本研究では、ある一つの製品についてのマニュアルから文を収集する。このマニュアルには、章立てで分割すると 3,000 以上の文書が存在し、その文書長は約 8,000 文から 1 文までさまざまである。係り受け付与対象の文書は無作為に選択するが、作業を円滑に行うため、200 文程度の大きさの文書を優先的に選んで用いる。選んだ文書に対して文分割を行う。文分割は、改行文字や「。」等を文区切りとみなすルールにより行う。この段階で、章題やヘッダー・フッターなどを係り受け付与対象から外すため、「。」で終わらない文を取り除く。加えて、丸括弧は、任意の場所に挿入され、係り受け解析での対処が非常に困難である [7] ため、本稿では解析の対象外とし、丸括弧を含む文は取り除く。

(2) 係り受けテンプレートの作成　本研究では、係り受けを図 1 の左上および左下のような形式のテキストファイルで表す。本稿ではこの形式を文節スライド形式と呼ぶこととする。文節スライド形式で書かれた文は、単語区切り、文節区切り、係り受けの情報を持つ。一行が一文節を表す。各行には、文節内の単語がスペース区切りで記載される。半角ハイフンの左側に内容語が、右側に機能語が置かれる。係り受け関係はインデント構造によって表される。文節の係り先は、その文節よりも一段浅いインデントの文節のうち最も近いものである。係り受けにラベルを付与する場合は、各行の末尾に半角ハイフンに続けてラベルを記述すればよい。

テンプレートの作成方法は次のとおりである。まず文の形態素解析を行う。形態素解析には MeCab [1] を用いる。辞書には、IPA 辞書と、マニュアルの製品に関する専門用語の辞書を用いる。ここで得た形態素

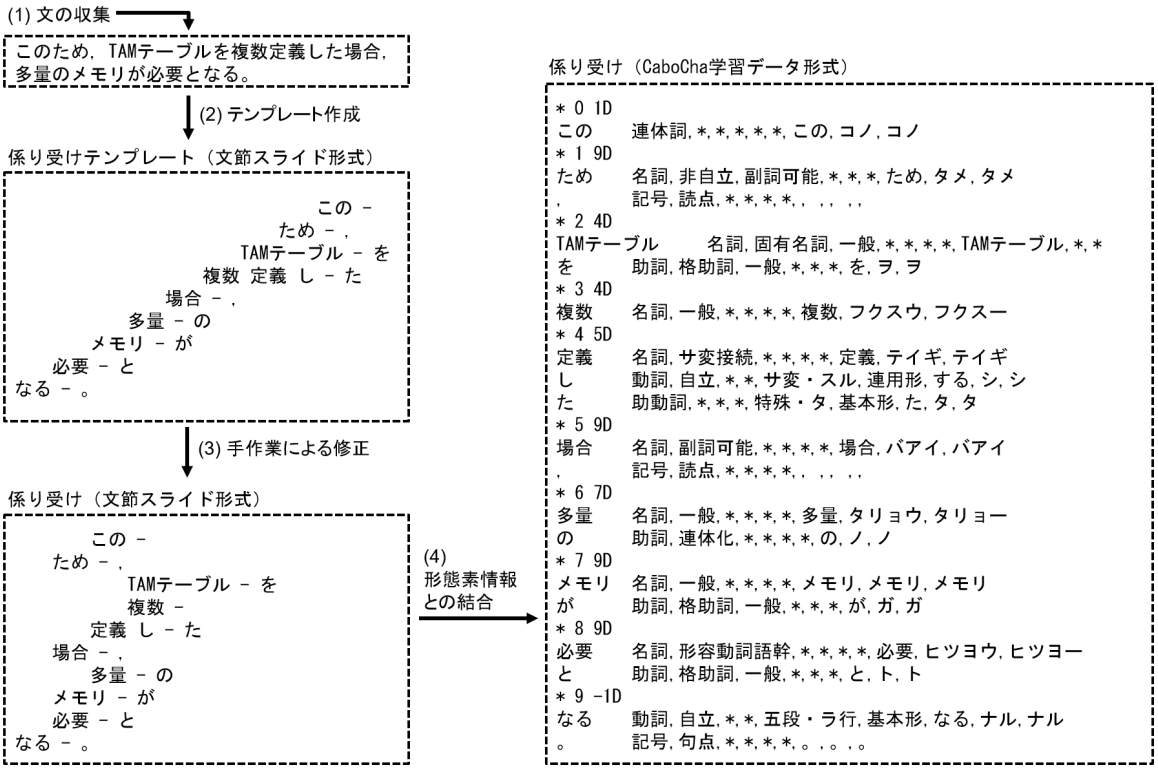

図 1: 係り受けコーパス作成手順と例

情報は (4) でも用いるため別途保存しておく。次に文節の区切りを仮決定する。区切りは、助詞、助動詞や記号の次で文節を区切るなどの、品詞に基づくルールによって決定する。最後に係り受けを仮決定する。テンプレートでは、すべての文節が直後の文節に係るとする。

(3) 手作業による修正　テンプレートを手作業で修正し、図 1 左下のような文節係り受け情報とする。文節の区切りと係り受けが修正対象であり、単語区切りは修正せず形態素解析結果をそのまま採用するものとする。文節スライド形式では、修正作業はすべて一つのテキストファイルの編集によって行うことができる。係り受けの修正はインデントの変更だけで行うことができる。

(4) 形態素情報との結合　最後に、修正された係り受け情報を、元の文の形態素情報と組み合わせて、図 1 の右のような形式に変換する。

2.3　複数の作業者による係り受け付与

係り受け付与は、同一の文に対して二名以上の作業者が行う。図 2 に作業者が二名である場合の作業の流れを示す。まず各々の作業者が係り受けを付与する。この結果を v0 データとする。次に、他の作業者の v0 データと比較して、差分を確認しながら自身の v0 データを修正する。修正後のデータを v1 データとする。ここまでの作業は各作業者が独立に行う。そして、互いの v1 データを比較し、差分を確認しながら、v2 データを作成する。作業者は、製品知識及び係り受け付与作業経験を持たない非専門家である。

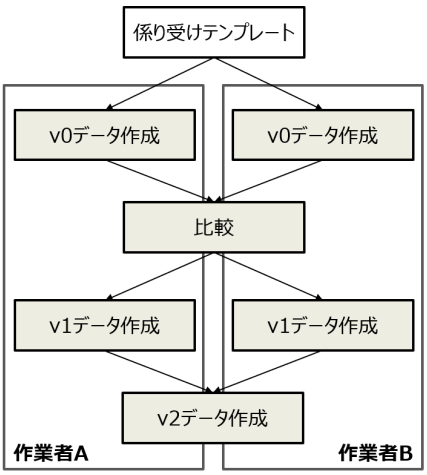

図 2: 二名の作業者による係り受け付与の流れ

表 1: 係り受けコーパスの統計情報

種類	ジャンル	記事数	文数	文節数	語数	平均文書長	平均文長	平均文節長
京都大学テキストコーパス [7]	新聞	2,927	38,400	372,130	972,894	13 文	9.7 文節	2.6 語
KNB コーパス [8]	ブログ	249	4,186	27,792	66,952	17 文	6.6 文節	2.4 語
マニュアルコーパス（試作）	技術文書	9	1,258	9,475	24,001	140 文	7.5 文節	2.5 語

表 2: 品詞分布

品詞	マニュアルコーパス	京大コーパス	KNB コーパス
名詞	41.6%	41.9%	31.4%
助詞	27.1%	26.3%	27.2%
記号	12.6%	11.2%	13.0%
動詞	11.6%	10.8%	12.6%
助動詞	3.4%	6.0%	8.9%
副詞	**0.4%**	1.0%	2.6%
形容詞	**0.4%**	0.9%	2.0%
接頭詞	0.7%	0.8%	0.4%
連体詞	0.8%	0.6%	0.8%
接続詞	**1.4%**	0.4%	0.8%

表 3: 作業者間一致率

	文節区切り		係り受け	
	文節	文	文節	文
v0 データ	0.89	0.84	0.80	0.62
v1 データ	0.97	0.96	0.94	0.83

3 結果

3.1 試作したコーパス

表 1 に試作したソフトウェアマニュアル係り受けコーパス（以下、マニュアルコーパス）と、既存の係り受けコーパスの統計情報を示す。マニュアルコーパスにおいても、平均文長は 7.5 文節、平均文節長は 2.5 語と、新聞記事やブログ記事を用いたコーパスの文章と同程度であった。

係り受け付与作業のペースは、v0 は 1 時間あたり約 90 文、v1 は 1 時間あたり約 180 文であった。

表 2 に品詞の分布を示す。京大コーパスと比べて倍以上の差がある品詞に注目すると、マニュアルコーパスには、副詞や形容詞が少なく、接続詞が多いという特徴があった。マニュアル文章の特徴である、副詞や形容詞で表されるような抽象的な情報は書かれない点や、手順の説明が多く並列が多用されている点が表れていると考える。

3.2 作業者間一致

表 3 に作業者二名の間の一致率を示す。v0 は作業者が独立に付与した結果の一致率であり、v1 はもう一名の v0 データとの差分を確認しながら自身の v0 データを修正したものの一致率である。文節区切りと係り受けについて、それぞれ文節単位での一致率と文単位での一致率を示している。文節単位での一致率は、文節区切りについては「一致した文節の数÷文節の数」、係り受けについては、係り先のない最後の文節を除いた「係り先の文節番号が一致した文節の数÷文節の数」である。ただし、係り受けについては、各作業者の文節区切りの違いによって分母の文節の数が異なる場合があるため、各作業者が相手の作業者の付与結果を正解とした場合の自身の付与結果の正解率を計算し、二名の正解率の調和平均を一致率とした。文単位での一致率は、各文について、その文内のすべての付与結果が一致する場合に一致、一つでも不一致がある場合に不一致とした。

表 3 のとおり、v0 データでは文節区切りは 84%、係り受けは 62%の文しか一致しなかったが、修正後の v1 データでは 96%、83%の文が完全に一致するようになった。つまり、v1 データ作成の工程により、文節区切りでは 12%、係り受けでは 21%の文において不一致が解決した。ここで解決した不一致は、v0 データ作成時の判断ミスが原因であったと考える。

v1 データの不一致の原因を文ごとに分類した。分類結果を表 4 に示す。16%が v1 作成時には修正しきれず、v2 データ作成時に明らかになった判断ミスであった。12%がマニュアル文章特有の文形であるために、作業時には係り受け判断基準が定められておらず、判断が一致しなかった文であった。このようなマニュアル特有の文の係り受けについては、次節で述べるように追加の基準を設定した。32%が製品知識の不足により係り受けが判断できない文であった。たとえば、

(1) 物理エリア単位の排他制御と UAP スケジュールは，データベースの場合と同様である。

という文では、「物理エリア単位の」の係り先には、「排他制御（と）」と「UAP スケジュール（は）」の二通りが考えられる。どちらが正しいかを判断するには、“UAP スケジュールが物理エリア単位で設定されるものであるか否か” という製品知識が必要となる。残りの不一致の 40%は、作業者に係り受け基準の知識が十分でないため、係り受けの判断に迷う文であった。

判断ミスによる不一致は修正により、マニュアル文章特有の文による不一致は基準の追加により、それぞれ対処が可能である。このように対処を行った後の最終的な作業者間一致率は 88%（=0.83+0.027+0.020）となった。

表 4: v1 データにおける作業者間の不一致の分類

不一致の原因	不一致文に占める割合	全体に占める割合
判断ミス	16%	2.7%
マニュアル文章特有の形	12%	2.0%
製品知識不足	32%	5.4%
係り受け基準の知識不足	40%	6.8%

形態素解析における単語区切りの誤りは若干数見られたが、正しい文節区切りを妨げる誤りは無かった。

3.3 追加した係り受け判断基準

マニュアルコーパス作成にあたり、以下の判断基準を新たに設定した。係り受けについては、係り元を下線、係り先を太字として、例文中に表す。

指示を表す「こと」で終わる文

(2) この句の USE 指定は，後続の DML で同一ページを再利用することが多いときに指定すること。

(3) プログラム間通信時の 出力内容については，表 17 － 19 を参照のこと。

指示を表す「～は、～（する | の）こと」という形の文は、「こと」を名詞と捉えて独立の文節とし、「～は」の係り先を「こと」とする。「こと」を終助詞として考え、手前の句とまとめて一文節と考えることもできるが、前段の形態素解析で用いた IPA 辞書では「こと」は名詞となっている。そのため今回は独立の文節を作る名詞として取り扱うが、議論の余地がある。

記号の定義を説明する文

(4) ○：排他モードが**遷移する。**

(5) タイプ３：レコード型のアクセス目的が RETRIEVE と UPDATE の**混在である。**

記号や標識の定義を説明する文では、説明対象の語句は末尾の述語に係るとする。

3.4 係り受け解析による評価

試作したマニュアルコーパスを利用して、文節区切りと係り受け解析の教師あり学習による評価を行った。試作したコーパスを分割し、全 1,258 文中の 1,058 文を学習用、200 文を評価用とした。解析器には CaboCha を用いた [6]。評価結果を表 5 に示す。現段階では評価用データの規模が小さいため精度はあくまで参考値であるが、コーパスの規模が大きくなるほど、解析の精度が向上する傾向が見られた。

表 5: 文節区切りおよび係り受け解析の精度

学習データ	文数	文節区切り	係り受け解析
マニュアルコーパス	215	0.74	0.59
〃	645	0.81	0.68
〃	1,058	0.83	0.71
京大コーパス	30,791	0.96	0.84
KNB コーパス	3,328	0.97	0.84

4 おわりに

技術文書に適した係り受け解析モデルの構築を目的として、製品マニュアルを用いた係り受けコーパス構築の試作を行った。その結果、1,258 文の係り受け情報を作業者間一致率 88%で付与することができ、製品知識及び係り受け付与作業経験を持たない非専門家でも十分な質のコーパス構築ができる見込みを得た。今後は、引き続き係り受け情報の付与を進め、10,000 文規模のソフトウェアマニュアル係り受けコーパスの構築及び係り受け解析モデルの作成を行う。

参考文献

[1] MeCab: Yet Another Part-of-Speech and Morphological Analyzer.

[2] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Asuharu Den. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, Vol. 48, pp. 345–371, 2014.

[3] Kohsuke Yanai, Misa Sato, Toshihiko Yanase, Kenzo Kurotsuchi, Yuta Koreeda, and Yoshiki Niwa. Struap: A tool for bundling linguistic trees through structure-based abstract pattern. In *Proceedings of EMNLP 2017 System Demonstrations*, pp. 31–36, 2017.

[4] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel Slicing : Scalable Online Training with Conjunctive Features. In *COLING2010(oral)*, pp. 1245–1253, 2010.

[5] 前川喜久雄, 籠宮隆之, 小磯花絵, 小椋秀樹, 菊池英明. 日本語話し言葉コーパスの設計. 音声研究, Vol. 4, No. 1, pp. 51–61, 2000.

[6] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[7] 黒橋禎夫, 長尾真. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 30 回年次大会, pp. 115–118, 1997.

[8] 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明. 構文・照応・評価情報つきブログコーパスの構築. 自然言語処理, Vol. 18, No. 2, pp. 175—-201, 2011.