異なる品詞体系に基づいて付与された言語コーパスの品詞タグ情報を再利用する

脇川浩和 *1　乾健太郎 *1 *2
*1 九州工業大学情報工学部知能情報工学科
*2 科学技術振興事業団さきがけ研究21
{h_waki,inui}@pluto.ai.kyutech.ac.jp

1 はじめに

自然言語処理の分野では近年，言語コーパスを，言語知識の学習や統計情報の獲得などに利用する研究が盛んである．一般に，コーパスに基づく自然言語処理システムの性能は言語データの量に依存する．そのため，このようなシステムをあつかった研究では，できるだけ大規模なコーパスを利用できることが望ましい．しかし，大規模なコーパスの作成には多くの時間と手間を要するため，十分な量のコーパスを確保するのは容易でない．そこで複数のコーパスを共有したり，再利用したりすることが求められている．しかしながら，コーパスによって品詞体系が異なり，それにともなう単語境界の認定基準も異なる場合が多く，コーパスの共有には何らかの工夫が必要である．

このような背景から本研究では，形態素・構文情報つきコーパスを再利用するために，既存のコーパスの品詞・構文タグ（ソース側タグ）を別の品詞体系に基づく品詞・構文タグ（ターゲット側タグ）に変換するアルゴリズムを提案する．本手法では，ターゲット側品詞体系に基づく文法・辞書でコーパスを形態素・構文解析することによって半自動的にタグ付けを行う．ただし，単純な解析方法では大量の曖昧性が残ってしまう．この問題を解消するために，本手法ではソース側タグ情報を最大限に利用し，ターゲット側の曖昧性を削減する．

2 品詞体系変換アルゴリズム

本手法では，ターゲット側品詞体系に基づく文法・辞書でコーパスを形態素・構文解析することによって半自動的にターゲット側のタグ付けを行う．ターゲット側では，品詞タグの他にこの解析には以下の資源を用いる．

- ターゲット側文法: ターゲット側品詞体系に基づく文節内文法
- ターゲット側辞書: ターゲット側品詞体系に基づく辞書
- ターゲット側連接表: 隣接する品詞間の連接可能性に関する制約
- ターゲット側係り受け表: 文節間の係り受け可能性に関する制約

ただし，これらの資源を使うだけでは，通常の形態素・構文解析と同じ作業になり，大量の曖昧性が残ってしまう．そこで，対象とするコーパスに以下のソース側タグ情報が与えられているものと仮定し，それらを利用することによってターゲット側の曖昧性を削減する．

- ソース側単語境界: ソース側品詞体系に基づく単語の境界
- ソース側品詞タグ: 各単語に付与されているソース側品詞タグ
- ソース側文節境界: ソース側の文節認定基準に基づく文節の境界
- ソース側係り受け情報: 文節間の係り受け関係

これらの情報はいずれも，EDR日本語コーパス [6] や京大コーパス [2]，ATRコーパス [4] といった既存の構文木つきコーパスから抽出できる情報であり，ソース側情報として仮定するのは自然だと考えられる．

本手法は次の手順からなる．

1. 前処理： コーパスから品詞変換表を半自動的に作成する
2. 文節内処理： 文節ごとに次の処理を行う
   2.1 品詞変換表を適用し，ターゲット側品詞タグ列の候補を生成する
   2.2 得られた品詞タグ列の候補をターゲット側文法で構文解析し，曖昧性を削減する
3. 文節間処理： 文節内処理で曖昧性が残った文節に対し，文節間の制約を適用することによって，曖昧性をさらに削減する

以下，それぞれの処理の概要を述べる．

2.1 前処理

前処理では，二つの品詞体系間の品詞変換表を作成する．品詞変換表は，

（ソース側品詞） → （ターゲット側品詞の列）

という形の品詞変換規則の集合である．たとえば，「広がりつつ」という文字列は，京大コーパスの品詞体系（益岡文法品詞体系）では，

広がり（子音動詞ラ行基本連用形）
つつ（接続助詞）

と解析されるが，EDR辞書の品詞体系では，

広が（ラ行五段動詞語幹）
り（ラ行五段活用語尾・五段連用形）
つつ（動詞連用形後接語・接続助詞）

と解析される．この対応からは以下のような品詞変換規則が得られる．

子音動詞ラ行基本連用形 →
　ラ行五段動詞語幹 ラ行五段活用語尾・五段連用形
接続助詞 → ラ行五段活用語尾・五段連用形

このような変換規則は，ソース側品詞タグとターゲット側品詞タグがともに付与されている訓練用コーパスがあれば，そこから自動的に抽出することができる．そのような訓練コーパスが手に入らない場合は，以下の手順で半自動的に収集する．

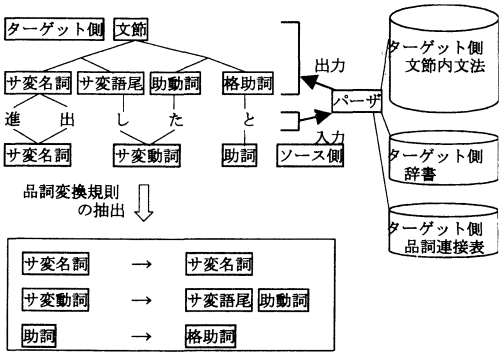

図 1: 品詞変換表の作成

2.1.1 品詞変換規則の候補の収集

ソース側のすべての品詞について品詞変換規則を漏れなく作成する必要がある．そこで，まずソース側の各品詞について，その品詞を含む十分な数の文節を用意する．つぎに，集めた文節をターゲット側の文節内文法・辞書・連接表を用いて解析する．解析で得られるターゲット側品詞タグ候補のうち，ソース側品詞タグと単語境界が矛盾しないものから，図1に示すように品詞変換規則の候補を生成する．ここまでの作業は人手を要しない．

2.1.2 品詞変換規則の洗練

上の作業では，ターゲット側の解析で曖昧性が生じるため，次のような誤った対応関係も品詞変換規則として収集されてしまう．

ナノ形容詞（ストレートな）→
普通名詞（スト）普通名詞（レート）助動詞（な）

これらの不適格な規則は，基本的には人手で取り除くしかない．3節で述べる実験では，対応関係の頻度情報や「名詞」「助詞」といった品詞の大分類の情報を手がかりにして，人手によって規則集合を洗練した．

2.2 文節内処理

2.2.1 品詞変換表の適用

このようにして得られた品詞変換規則は，ソース側の品詞を非終端記号として左辺に持ち，ターゲット側の品詞（列）を前終端記号として右辺に持つ文法規則と見なすことができる．このことを利用すれば，図2 に示したように，品詞変換表をベースとする文法を用いて構文解析することにより，コーパスに品詞変換規則の制約を効率的に適用することができる．

図2 のように，この構文解析の入力は，ソース側単語境界とソース側品詞タグの情報を付与した一文節の文字列である．パーザは，これらソース側の制約に無矛盾な構文木だけを出力する．解析に当たっては，ターゲット

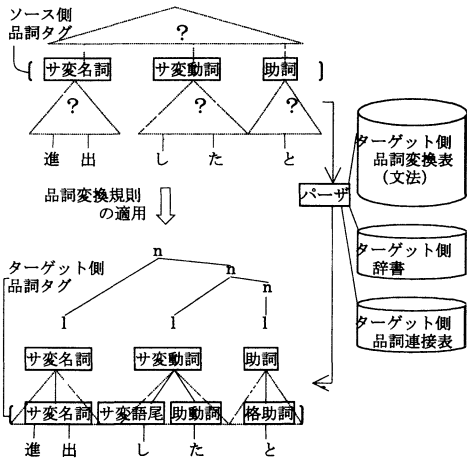

図 2: 品詞変換表の適用

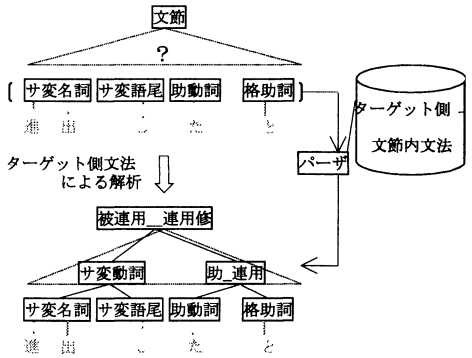

図 3: ターゲット側文法による解析

側辞書の他に，ターゲット側品詞連接表の制約を適用することによって解析結果の曖昧性の抑制する[1]

2.2.2 ターゲット側文法による解析

このフェーズでは，ターゲット側文節内文法の制約を適用することによって，2.2.1の処理で残る曖昧性を削減する．具体的には，2.2.1の処理で得られるターゲット側品詞列の各候補とターゲット側文節境界を入力として，ターゲット側文節内文法による構文解析を行う（図3）．

この過程で，文節内文法に合わない品詞列は棄却される．一方，文法に合った品詞列は文節にまとめ上げられる．文節内文法をうまく設計しておけば，図3のように，文節の属性を表すラベルを文節の節点に割り当てることも可能である．このラベルは次の文節間処理で利用する．

入力にターゲット側文節境界の制約を入れるのは，文節境界の制約が曖昧性解消に有効であることが経験的に

[1]我々が実験で用いた東工大で開発された MSLR パーザ [3] は，文脈自由文法と品詞連接表の制約を同時に適用して形態的・構文的曖昧性を抑制することができる．また，同パーザは，入力文に部分的な構文構造が与えられると，それに無矛盾な構文木だけを出力することができる．

明らかになっているためである．ただし，ソース側とターゲット側で文節境界の認定基準が異なる場合がある．たとえば，「広がりつつある」という文字列には，全体で1文節とする解釈と，「広がりつつ／ある」のように2文節とする解釈が考えられる．このような場合，上述の文節内文法による解析は受理されない．そこで，上述の解析で受理されなかった文節列については，ターゲット側文節境界の制約を取り除いて，解析を再度行う．これによって，ソース側文節境界に比べてターゲット側文節境界が細かい場合は，扱うことができる．ただし，逆の場合は現在のアルゴリズムでは扱えないので，何らかの工夫が必要である．

2.3 文節間処理

ターゲット側の曖昧性は，上で述べた文節内の制約だけでは解消できないものも多いと予想される．たとえば，「太郎／と」のような文節では，「と」が格助詞なのか並立助詞なのか，文節内の情報を参照するだけでは決定できない．ところが，助詞「と」の場合について言えば，それを含む文節が用言に係っていれば格助詞，体言に係っていれば並立助詞であることがわかる．このように文節内処理で残る曖昧性の中には，文節間の係り受け情報を利用すれば解消できるものがある．

文節間処理では，文節境界をまたいで隣接する2つの品詞の連接可能性のチェック，および文節ラベルを用いた係り受け可能性のチェックを行う．

3 実験

提案したアルゴリズムの有効性を確認するための事例研究として，以下のセッティングのもとに実験を行った．

- ソース側：
  - コーパス：京大コーパス 10,697 文（105,885 文節）
  - 品詞体系：益岡文法に基づく品詞体系（品詞数 416 個）
- ターゲット側：
  - 品詞体系：EDR 日本語単語辞書に基づく細品詞体系（品詞数 621 個）[2]
  - 文節内文法：東工大研究グループが開発した文法 [6] を利用．品詞列を文節単位にまとめ上げ，各文節に係り属性（連用／連体）と受け属性（連用／連体／無）の組を表す非終端記号を割り当てる．
  - 辞書：EDR 日本語単語辞書を拡張して使用
- 解析器：東工大 MSLR パーザ [3]

3.1 品詞対応表の作成と辞書の拡張

まず，ソース側の各品詞について，それを含む文節を15個ずつ用意した．これらの文節に対し2.1.1で述べた処理を行ったところ，約 4,000 個の品詞変換規則の候補が得られた（表1参照）．これはソース側品詞数の約10倍である．規則の候補がこのように多くなるのは，次の例のように，形態素・構文解析の段階で，ソース側の単語を細かく分割してしまう傾向が強いためである．

[2]ここで用いる細品詞体系は，EDR 辞書の品詞と左右連接属性の組み合わせを一意に特定するように定義された品詞ラベルの集合である．

表 1: ソース側品詞数と品詞対応規則数

自動獲得した規則候補数	3935 パターン
手作業による洗練後の規則数	631 パターン
助詞・副詞などの追加後の規則数	704 パターン

ナノ形容詞（ストレートな）→
　普通名詞（スト）普通名詞（レート）助動詞（な）
イ形容詞（女らしく）→
　普通名詞（女）助動詞（らし）助動詞（く）

このようにして得た規則候補を人手によって取捨選択したところ，600 余りの規則しか残らなかった．ただし，EDR 細品詞体系では，助詞や副詞など，一部の品詞については非常に細かい分類がなされているので，ソース側各品詞について 15 文節を用意するだけでは，得られる変換規則に漏れが生じることが作業の過程で明らかになった．そこで，EDR 辞書マニュアル [1] の記述を手がかりにして，規則の追加を行なった．追加した規則は 73 個であった．

辞書は，EDR 日本語単語辞書をもとに作成した．ただし，固有名詞をはじめとする名詞類の未登録単語については，ターゲット側とソース側で単語境界が一致している語をコーパスから抽出し，ターゲット側辞書に登録した．その他の未登録単語については，手作業で登録した．とくに表記の違い（ゆれ）によるものが多かった．

3.2 タグづけ

3.2.1 品詞変換表の効果

品詞変換表によって与えられる制約の効果を調べた．比較対象は，品詞変換表を用いず（すなわち，ソース側のタグ情報を利用せず），ターゲット側文節内文法による形態素・構文解析だけでタグづけを行う方法である．ただし，文節品詞変換表を用いない場合にも，ソース側とターゲット側で文節境界が一致するという制約を与えて解析を行った．結果は表2に示す通りである．

また，変換テーブル適用（品詞ラベル付与）後に，候補が一意に決定したもののみ文節内文法で解析を行った．その結果も合わせて表2に示す．表からもわかるとおり，単純な解析方法と比べて，解析結果が一意に決定する割合が，変換テーブルによる解析で 42%，その後の文節内文法による解析との比較では 34% 向上している．

3.2.2 文節内文法の効果

品詞変換表の適用後，文節内文法による構文解析を行ったところ，809 文節が文法により拒否され，曖昧性を取り除かれ一意に決まったものは 1184 文節であった．文法で拒否されたもののほとんどは文法が未対応のものであった．また，文節区切り制約なしの解析では，ソース側の文節認定基準よりターゲット側の認定単位基準が細かい場合の，文節内構造を拾い揚げることができた．

3.2.3 文節間処理の効果

文節内処理で曖昧性があった文節のうち，文節間処理によって一意に決定できた文節は，4,606 個で 13% に過ぎなかった．文節間処理があまり効果的でなかった理由としては，今回の実験で用いた文法の文節ラベルが荒

表 2: 実験結果

変換規則を用いない場合		変換規則を用いた場合						
		規則のみ		規則適用後文法で解析			文節間の制約		
一意に決定	15,946	(u)	68,387	(u) から	58,201	(u')	(u') から	48,845	53,451
(u)nique				(a) から	1,184	59,385	(a') から	4,606	50%
曖昧性あり	83,693	(a)	27,574	(u) から	9,661	(a')	(a') から	25,225	25,225
(a)mbig				(a) から	26,106	35,767			24%
解析失敗	6,246	(r)	9,924	(u) から	9,661	(r')	(u') から	10,540	27,209
(r)ejected				(a) から	26,106		(a') から	5,936	26%
				(r) から	9,924	10,733	(r') から	10,733	

すぎたため，係り受け関係の制約が効果的に働かなかったことが挙げられる．

文節間処理を施しても一意に決定できなかった文節については，助詞の曖昧性が最も多く 7 割強，次いで名詞句の構造の曖昧性が 1 割弱であった．ランダムにサンプリングしたデータを人手で調査した結果，助詞の曖昧性は，いずれも品詞の定義自体の曖昧性が原因になっていることがわかった．例えば，EDR 細品詞には「体言にのみ後接する格助詞の」と「格助詞に後接可能な格助詞の」があるが，前者は後者に包含されるので，体言に「の」が後接した場合，両者を区別することは不可能である．名詞句の構造の曖昧性については，複合名詞句内の構造の曖昧性が大部分を占めた．複合名詞句の構造は，提案したアルゴリズムで用いる制約からは一意に特定するのがほとんど不可能であるので，曖昧性を明示的に表現しない文法に変更する必要があると考えられる．

文節間処理で受理されなかった文節については，「名詞句＋判定詞」あるいは「名詞句＋読点」といった文節の文節内文法での扱いに問題があって，係り受け関係の制約を満たさなかったものが 7 割近くを占めていた．たとえば，「日本は中国に進出，」の「進出，」は，現在の文法では連体修飾のみを受ける文節として扱われるので，「日本は（係り属性：連用）」がそこに係ることができない．

このように，文節間処理で一意に決まらなかった文節について調査した結果，少なくとも今回の実験のセッティングに関するかぎり，品詞体系の問題と文節内文法の問題がもっとも重要であることがわかった．

3.2.4 一意に決定されたターゲット側タグの信頼性

最後に，本手法によって一意に決定されたターゲット側タグ情報の信頼性を確認する調査を行った．一意に決定された文節の中からランダムに 100 文節を抽出し，正しく解析されているかどうか人手で調べた結果，100 文節とも正しく解析されていることがわかった．小規模な調査ではあるが，この結果は，一意に決定されたターゲット側のタグが十分に信頼できることを示唆している．

4 関連研究

品詞体系の自動変換に関してはすでにいくつかの先行研究が見られる．

たとえば植木らは，品詞体系の違いは文節内構造にしか影響しないと考え，文法を品詞体系非依存部分 (文節間構造) と依存部分 (文節内構造) に分け，品詞体系により依存部分の文法のみを入れ替え，非依存部分に共通の文法を用いることで品詞体系の違いを吸収しようと試みている [6]．植木らの考え方は，本稿で述べた変換アルゴリズムともよく整合する．

また田代らは，本稿で述べた品詞変換表と似たような情報を訓練コーパスから抽出し，それに基づいて品詞体系を自動変換するアルゴリズムを提案している．品詞の対応パターンを曖昧性解消に用いるという考え方は我々のアルゴリズムも同じである．ただし，我々のアルゴリズムでは，ソース側の文節境界情報や係り受け情報，ターゲット側の文節内文法や係り受け制約，といった多様な制約を使う点が異なる．

5 おわりに

本稿では，形態素・構文解析器を用いて既存のコーパスのタグを異なる品詞体系に変換するアルゴリズムを提案した．品詞タグや係り受け情報など，ソース側の情報を最大限に利用することにより，単純に形態素・構文解析する場合に比べ，タグづけの曖昧性を大幅に削減できることを，一つのケーススタディを通して示した．ただし，本稿で提案したアルゴリズムのままでは，ソース側の単語・文節境界よりもターゲット側の単語・文節境界が荒い場合を扱うことができない．この点の拡張が今後の課題である．

謝辞

実験に当たっては，東京工業大学で開発された MSLR 構文解析器を利用させていただきました．同大学の田中穂積氏，白井清昭氏，植木正裕氏，橋本泰一氏に感謝いたします．

参考文献

[1] EDR. 電子化辞書仕様説明書 第 2 版. Technical report, 日本電子化辞書研究所,3 1995.

[2] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会予稿集, pp. 58–61, 1997.

[3] Li, Tanaka A method for integrating the connection constraints into an LR table. In *Proceedings of Natural Language Processing Pacific Rim Symposium '95*pp703–708,1995.

[4] 田中穂積 (東工大), 竹澤寿幸 (ATR), 衛藤純司 (ランゲージウェア). MSLR 法を考慮した音声認識用日本語文法. 情報処理学会研究報告 (音声言語処理研究会)15-25, pp. 145–150, 1997.

[5] 田代敏久, 森元. 形態素情報付きコーパスの再構築手法. 情報処理学会論文誌, Vol.37, No.1, pp.13–22,1 1996.

[6] 植木正裕, 白井清昭, 徳永健伸, 田中穂積. 構造つきコーパスの共有化に関する一考察. 情報処理学会研究報告 (98-NL-128)128-9, pp. 61–66, 1998.