形態素解析に依存しない日本語係り受け解析

吉橋健治　仁科喜久子
東京工業大学
knj4484@gmail.com　nishina.k.aa@m.titech.ac.jp

１．はじめに

従来、日本語データを処理し、有用なデータを抽出するアプリケーションシステムでは、1)形態素解析、2)文節のまとめ上げ、3)係り受け解析、4)情報抽出のような流れで処理が行われており、通常は、1)～3)の処理にはアプリケーションに依存しない汎用的なツールが用いられる。構文解析手法が品詞などの形態素情報を前提としており、前処理として形態素解析を必要とするので、このような処理順序が標準的に用いられている。

しかし、この手順にはいくつかの問題点が含まれている。まず、形態素解析では、以降の処理で必要とされている以上に詳細な解析をしており、非効率である。例えば、複合語の切り分けや、品詞細分類は、構文解析やアプリケーションで必ずしも必要というわけではない。また、形態素解析で用いられる品詞体系や解析結果が、最終的な情報抽出に適しているとは限らない。例えば、評判情報抽出において、体言と述語の組み合わせの一つとして「おいしさ$_{体言}$ が アップ$_{述語}$」という表現を簡単には抽出することができない。なぜなら、「おいし$_{形容詞}$ さ$_{接尾辞}$ が アップ$_{名詞}$」と解析されてしまうからである。最後に、文節のまとめ上げは比較的簡単なタスクであるのに、より困難なタスクである形態素解析の結果を利用しているのは合理的とは言えない。

これらの問題点が示唆していることは、言語処理システムの構成において、形態素解析は行わないか、それに代わる処理を目的に合わせて情報抽出の直前に行った方がよいということである。この考え方によれば、1) 文節のまとめ上げ、2) 係り受け解析、3) 形態素解析、4)情報抽出という処理順序が理想的ということになる。本稿では、このような言語処理の構成の中で、文節のまとめ上げまでは完成しているとし、形態素情報なしの文節の並びから係り受け解析を行う統計的手法を提案する。

２．提案手法

2.1 解析モデル

解析モデルには、相対的な係りやすさを考慮した日本語係り受け解析モデル[4]を用いた。この解析モデルは、手順がシンプルでありながら、解析精度が高いことが実験的に示されている[4]。解析手順は以下の(1)～(3)の通りである。

(1) 文末から文頭へと順に係り先を決める
(2) 非交差条件により係り先候補を絞る
(3) スコアが最大の候補を係り先とする

スコアは、係り側文節と係り先候補の素性ベクトル Φ、素性ベクトルの重みパラメータ w の内積 w・Φ とする。スコアが最大値をとる候補が複数存在する場合は、係り元に最も近い文節を選択する。

2.2 形態素情報を用いない素性

提案手法では、形態素情報の代わりに文節中の部分文字列を素性とする。具体的には、文節の前から 1 文字、2 文字、…、n 文字（PREFIX 文字列）、および後ろから 1 文字、2 文字、…、n 文字（SUFFIX 文字列）を抜き出したもののみを使うことにした。この際、記号や句読点なども文字として扱う。係り文節と受け文節で上記の部分文字列をそれぞれ 2n 個ずつ列挙し、係り PREFIX-受け PREFIX、係り SUFFIX-受け PREFIX、係り SUFFIX-受け SUFFIX の組み合わせで $3n^2$ 個の素性を作る。

部分文字列の他に、係り文節から受け文節までに含まれる鍵括弧と丸括弧のパターンで 2 個の素性を作り、係り文節と受け文節の間の文節について、a)句点あり読点なし、b)読点あり、c)句読点なしの 3 個の素性を作る。また、係り文節のPREFIX文字列と受け文節のSUFFIX文字列を文節間距離と組み合わせて 2n個の素性を作る。ここで、文節間距離として、距離の数値そのものではなく、A)１～２、B)２～５、C)６以上の３値とした[1]。

2.3 パラメータの推定

本研究では、訓練データ数が膨大になるため、パ

[1] 従来の研究ではA)１、B)２～５、C)6以上の３値であったが、正しい係り受け距離が１で解析結果が２となる誤りパターンとその逆の誤りパターンが非常に多いため、A を１～２に変更した。いくつかの既存の手法で、この変更を適用して実験してみると、精度が向上することが確認された。

ラメータ w の学習には、オンライン型の学習ができ、かつアルゴリズムがシンプルなパーセプトロンを用いた。図１にそのアルゴリズムを示す。

```
foreach sentence in 訓練データ
  for i = 1…T
    現在のパラメータで sentence を解析
    if 解析結果の構文木 P が間違っている
      w = w + Φg(P) − Φg(Pc(sentence))
```

図１　パラメータ学習アルゴリズム

T は反復回数、Φ_gは係り受け素性ベクトルの文全体にわたる和、P_cは訓練データの構文木である。

３．評価実験

提案手法の解析精度について、以下のデータを用いて評価実験を行った。

訓練データ：京都テキストコーパス[2]（一般記事１～13 日、社説１～９月分）、および毎日新聞 CD-ROM93、94、96 年分（CaboCha[3]により文節まとめ上げと係り受けの情報を付与した）

テストデータ：京都テキストコーパス（一般記事 14～17 日、社説 10~12 月分）

部分文字列の文字数nと学習での反復回数Tを変化させて精度[4]を測定した。訓練データ数約 210 万文の時点での精度を表１に示す。精度は、反復回数にはよらず、文字数が多い方が高いことが分かる。

表１　提案手法の係り受け正解率[%]

文字数n		1	2	3	4	5
反復回数T	2	80.6	88.0	89.3	89.5	89.7
	4	81.0	87.8	89.0	89.6	89.5
	8	81.4	88.1	89.3	89.6	89.8
	16	80.5	87.9	89.3	89.6	89.6
	32	81.1	88.0	89.4	89.6	89.7

４．従来の構文解析手法との比較

提案手法と従来の解析モデル（KNP[1]、後方文脈モデル[2]、チャンキングモデル[3]、相対モデル[4]、複数格制約モデル[5]、並列・格構造統合確率モデル[6]、線形時間モデル[7]）の精度[5]の比較を表２に示す。訓練データ欄のKCは京都テキストコーパスを表す。素性欄において、Bは注目する係り文節と受け文節のみの素性、C は格フレームに関する情報、D は動的素性（または注目文節の前後の文節の素性）の利用を意味する。提案手法は、訓練データを肥大化させることなく、シンプルな素性設計で高精度の解析を実現していることが分かる。また、既存手法と比較して提案手法には、(1)文法的、語彙的なルールを区別する必要がない、(2)品詞体系や辞書を作成する必要がない、(3)品詞体系や未知語の有無に影響されない、(4)訓練データに形態素タグを付与するコストが削減できるなどの利点がある。

表２　提案手法と先行研究の比較

モデル	訓練データ	素性	精度%
KNP	なし	B	*86.7
後方文脈	KC約8千文	B	87.93
チャンキング	KC約8千文	BD	89.29
相対	KC約2万5千文	BD	91.37
複数格制約	KC+新聞30年	BCD	91.25
並列・格統合	KC+Web5億文	BC	*87.4
線形時間	KC約8千文	BD	89.56
提案手法	KC+新聞3年(=約210万文)	B	89.76 *88.3

５．今後の課題

形態素情報を用いない日本語解析の研究は始まったばかりである。今後は、素性、解析モデル、学習方法など様々な点から精度向上を目指していきたい。

参考文献

[1]黒橋禎夫, 長尾眞. 並列構造の検出に基づく長い日本語の構文解析, 自然言語処理, vol 1, No.1, pp.35-58, 1994

[2]内元清貴, 村田真樹, 関根聡, 井佐原均. 後方文脈を考慮した係り受けモデル, 自然言語処理, vol 7, No.5, pp.3-17, 2000

[3]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

[4]工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析モデル, 情報処理学会論文誌, vol 46, No.4, pp.1082-1092, 2005

[5]阿辺川武, 奥村学. 共起情報及び複数格の組み合わせを考慮した係り受け解析. 自然言語処理, vol 13, No.2, pp.43-62, 2006

[6]河原大輔, 黒橋禎夫. 大規模語彙的知識に基づく構文・並列・格構造解析の統合的確率モデル, 言語処理学会第 13 回年次大会, pp.506-509, 2007

[7]颯々野学. 日本語係り受け解析の線形時間アルゴリズム, 自然言語処理, vol 14, No.1, pp.3-18, 2007

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html Version 4 を用いた。

[3] http://chasen.org/~taku/software/cabocha/ IPA-dep-A+P.model を用いた。

[4] 精度は、文末以外の文節の中で正しい係り先が得られた比率である。

[5] 精度欄の＊は文末から２番目の文節を評価に含めていないことを示す。