系列ラベリングによる準話し言葉の日本語係り受け解析

今村 賢治
日本電信電話株式会社　NTT サイバースペース研究所
imamura.kenji@lab.ntt.co.jp

1　はじめに

係り受け解析は、日本語における文節間の依存構造を解析する、構文解析の一種である。現在、このような解析器として、南瓜 (工藤・松本, 2002)、KNP (Kurohashi and Nagao, 1994) が広く普及している。これらは新聞記事を対象として作成された解析器である。

一方、近年、Web2.0 などの普及に伴い、エンドユーザが直接発信する文書がインターネット上で増加してきている。このような文書は統制された書き言葉ではなく、フィラーや顔文字などが含まれた、比較的話し言葉に近い文書であるため、従来の新聞記事を対象とした係り受け解析器では、解析が難しいと考えられる。

一般的には、構文解析は構文構造を 1 つの木で表し、木全体の尤度を最大化する構造を求めることにより解析を行う場合が多い (Charniak, 2000; 内元他, 1999; 工藤・松本, 2004) 。そのためフィラーなど、木構造では表しにくい要素の扱いが困難で、従来は強制的に他の要素に係るようにしていた。

本稿で提案する方式は、南瓜[1]の方式をベースとし、CRF に基づく系列ラベリングを用いて準話し言葉に対応できるように一般化したものである。具体的には、「着目文節がその後方 N 文節のうち、どの文節に係るか、あるいは係らないか」ラベルを、系列ラベリングによって付与し、それを再帰的に行うことにより係り受け解析を行う。系列ラベリングの際、「自分自身に係る（以下、これを自己係りと呼ぶ）」ラベルを含めることにより、フィラー等は自己係りとして解析される。つまり、文全体が 1 つの木である必要がなく、従来法に比べ、話し言葉に適した柔軟な解析が可能となる。

2　方式

2.1　チャンキングの段階適用による係り受け解析（南瓜）

南瓜は、決定的な解析を行う shift-reduce パーサの一種で、「着目文節が直後の文節に係るか否か（またはどのタイプでかかるか）」を、SVM に基づく分類器により判定する。1 回の判定では長距離の係り受けを考慮できないため、係り先が決定し、かつ他の文節が係らないと保証された文節を削除する。そして、縮退した文節列の係り受け判定を繰り返すことにより係り受け解析を行う。つまり、木全体の尤度を最大化するのではなく、局所的な係り受け解析を決定的に行い、これを再帰的に繰り返している。

南瓜は直後の文節に係るかを判定するものなので、原則、係り元、係り先の 2 文節の素性を考慮している。複数の文節を同時に考慮した判定を行うため、間素性と動的素性を導入し、削除された文節や、解析済み文節の素性を前後の文節に引継いでいる。しかし、自己係り文節は、考慮されていないため、必ず自分より後方の文節に係るよう、設計されている。

2.2　系列ラベリング

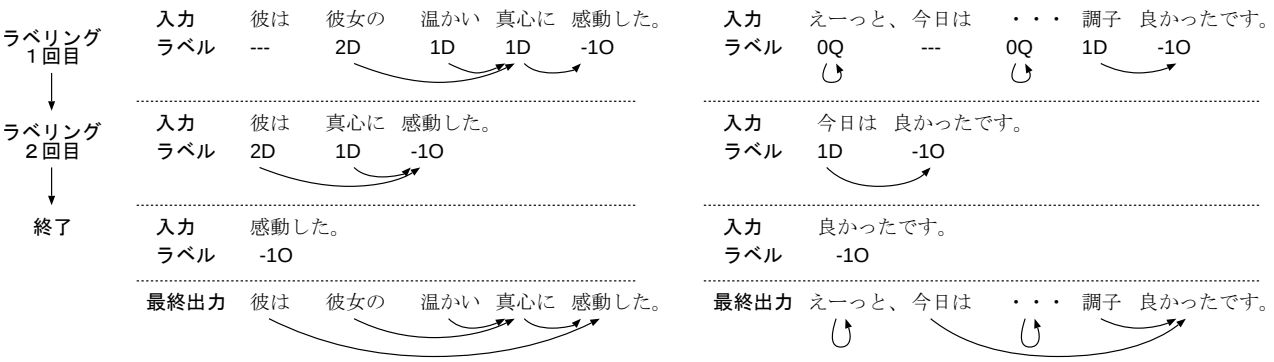
系列ラベリングは、入力列に対して適切なラベル列を推定し付与するタスクである。自然言語処理においては、英語の品詞タグ付けや固有表現抽出などに適用されており、隠れマルコフモデル (HMM) や、条件付確率場 (CRF; (Lafferty et al., 2001)) に基づくモデルが使用されている。本稿では、Linear-chain CRF に基づく系列ラベリングを使用する。

系列ラベリングでは、入力列に対してどのようなラベル（タグ）を付与するかは特に制約はなく、学習データの設計者が自由に設定できる。

2.3　提案方式

本稿で提案する係り受け解析は、南瓜方式を一般化したものである。係り先を決定する際、直後の文節に限るのではなく、着目文節（係り元）とその後方 N 文節（この範囲をウィンドウと呼び、N をウィンドウサイズと呼ぶ）のどこに係るか、あるいはウィンドウ内の文節には係らないのかを、系列ラベリングにより判定する。入力は、形態素解析済み文節列である。本方式の解析は以下の手順で実行される。

[1] http://www.chasen.org/~taku/software/cabocha/



(a) 通常の係り受けの場合　　(b) 自己係りを含む係り受けの場合

図 1: 係り受け解析例（ウィンドウサイズ 2 による解析）

表 1: 系列ラベリングで付与するラベル一覧 (ウィンドウサイズ 2 の場合)

ラベル	意味
—	ウィンドウ内に係り先なし
0Q	自己係り
1D	直後の文節に係る
2D	2 つ先の文節に係る
-1O	木構造のトップの文節

1. 文節から素性を抽出する
2. 素性を用いて系列ラベリングを行う
3. ラベルを解釈し、実際の係り先文節を決定する
4. 他の文節が係らないと保証された文節を削除し、文を縮退させる
5. 残った文節が 1 文節以下なら終了。そうでなければ、ステップ 1 に戻る

本方式による書き言葉の係り受け解析例を図 1(a) に示す。

ステップ 1, 2 では、通常の系列ラベリングと同様に、文節に対してラベルを付与する。ただし、係り受け解析用の系列ラベリングであるため、以下の特徴がある。

- 付与するラベルは、着目文節から見た係り先文節の相対位置を表すものである（表 1)。したがって、ラベルの種類は、ウィンドウサイズにより変化する。ウィンドウサイズを大きくとると、長距離の係り受けを 1 回のラベリングで解析することができるが、ラベルの種類が増加するため、データスパースネス問題が生じる。
- このラベルに自己係り（0Q）を含めておき、フィラー等、木構造にできない文節に対しても係り先を付与する。
- 特殊ラベルとして、「ウィンドウ内に係り先なし (—)」と、「木構造のトップ文節 (-1O)」を含む。“—” ラベルは、ウィンドウサイズが 2 の場合、3 つ先以降の文節に係ることを意味する。
- 着目文節のラベルを決定するための素性は、ウィンドウ内の文節の情報、および直前文節のラベルを使用する（図 2)。

ステップ 4 では、南瓜と同様に、「他の文節が係らない」と保証された文節を削除する。この際、係り受け非交差原則が適用される。たとえば、ある文節をまたぐ係り受けがある場合、その文節は他の文節が係らないので削除する。同様に “—” ラベルは N 個以上の文節をまたぐ係り受けであるので、その中に “—” ラベルを持つ文節がない場合、後方 N 文節を削除する。

以上を繰り返して文節列を短縮して、1 文節になったところで解析を終了する。図 1(a) の例では、延べ 2 回のラベリングで 1 文節になるため、終了する。

フィラーを含む文の場合、図 1(b) のように、系列ラベリングによって自己係りラベルが付与され、木構造とは独立した要素として解析される。このように、系列ラベリングのラベルとして自己係りを含めることにより、本方式はフィラーを含む文の解析が可能である。

3 実験

3.1 実験設定

コーパス　本実験では、2 種類のコーパスを使用した。まず、書き言葉のコーパスとして、京都テキストコーパス 4.0[2] を、準話し言葉コーパスとして、インターネット上のブログから収集した文（ブログコーパス）を使用した。コーパスサイズを表 2 に示す。なお、京

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

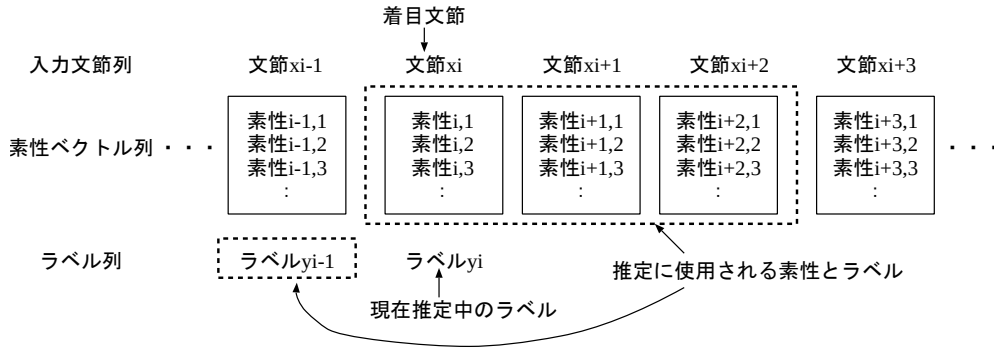


図 2: ウィンドウサイズが 2 のときに使用される素性

表 2: コーパスサイズ

コーパス	種類	文数	文節数
京大	学習	24,283	234,685
	テスト	9,284	89,874
ブログ	学習	18,163	106,177
	テスト	8,950	53,228

表 3: 使用した素性一覧

単独	主辞	見出し、品詞、品詞細分類、活用、活用形
	語形	見出し、品詞、品詞細分類、活用、活用形
	その他	開括弧有無、閉括弧有無、句読点有無、文頭か、文末か
組合せ		上記単独素性のそれぞれについて、着目文節とウィンドウ内の他文節で結合した素性

都コーパスは、学習用として一般記事 1 月 1–11 日、社説 1–8 月を、テスト用として一般記事 1 月 14–17 日、社説 10–12 月を使用している。ブログコーパスは、JUMAN[3] を用いて形態素解析し、京都コーパスと同様の係り受け情報を付与したものを使用した。

学習　本実験では、CRF++[4]を用いて学習を行った。素性は、多くの係り受け解析器 ((内元他, 1999; 工藤・松本, 2002) 等) で使用されているもののうち、文節固有のもの[5]を使用した。なお、CRF では素性の組み合わせは人手で決める必要があるので、今回は着目文節と、その後方文節の各素性を対にして組み合わせ素性を作成した。素性一覧を表 3 に示す。

評価方法　あらかじめ文節単位に分割された形態素列を入力として、係り受け正解率と文正解率（1 文について、すべての係り受けが正解したときのみ正解としてカウントする）を測定した。日本語の係り受け解析では、通常最終文節は係り先がないため、正解率算出から除外されるが、本方式の場合、最終文節が自己係りになることがあるため、最終文節も含めて係り受け正解率を算出した。

[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[4] http://www.chasen.org/~taku/software/CRF++/
[5] 南瓜における静的素性と同じ。

3.2　係り受け解析の精度

学習・テストコーパスを組み合わせて、係り受け解析を行った。提案方式のウィンドウサイズは 3 で固定した。また、比較のため、南瓜による係り受け正解率の測定も同時に行った。結果を表 4 に示す。

まず、京都コーパステストセットの係り受け正解率を見ると、南瓜が最もよい結果となった。これは、南瓜は 2 次の多項式カーネルを用いて素性の最適な組み合わせを自動的に発見し、解析を行うのに対して、本方式は素性の組み合わせのうち一部しか用いていないことが原因として考えられるが、今後検討が必要である。

提案方式は南瓜に比べ劣っていたが、京都コーパス + ブログモデルのように、異種コーパスを混合して訓練しても、京都コーパス単体のモデルに比べ、あまり精度は低下していない。

一方、ブログコーパステストセットの解析では、京都 + ブログモデルで正解率 84.59%と最もよい結果となった。これは、学習コーパスにブログが含まれているためもあるが、自己係りを解析できるようになった効果も大きい。実際、ブログコーパステストセットには、自己係りとなる文節は 3,089 個現れているが、そのうち 2,326 文節 (74.3%) を正しく解析できた。これは、コーパス追加に伴う係り受け正解率の向上のうち、約 6 割を占めている。

本方式は、自己係りを解析できることが特徴であるので、書き言葉の解析では劣るが、ブログのような準話し言葉では効果を発揮する。

表 4: 方式／テストセット別係り受け解析精度

テストコーパス	方式	学習コーパス（モデル）	係り受け正解率	文正解率
京都	提案方式 (ウィンドウサイズ 3)	京都	89.87% (80766 / 89874)	48.12% (4467 / 9284)
		京都 + ブログ	89.76% (80670 / 89874)	47.63% (4422 / 9284)
	南瓜	京都	**92.03**% (82714 / 89874)	**55.36**% (5140 / 9284)
ブログ	提案方式 (ウィンドウサイズ 3)	京都	77.19% (41083 / 53226)	41.41% (3706 / 8950)
		京都 + ブログ	**84.59**% (45022 / 53226)	**52.72**% (4718 / 8950)
	南瓜	京都	77.44% (41220 / 53226)	43.45% (3889 / 8950)

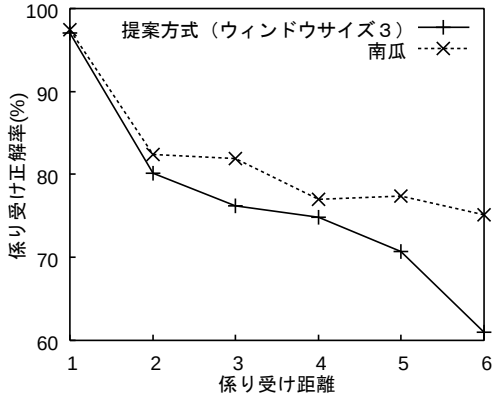


図 3: 係り受け距離別精度
(学習／テストともに京都コーパス)

3.3 係り受け距離別の精度

ウィンドウサイズを固定したときの係り受け距離別正解率を図 3 に示す。このグラフは、ウィンドウサイズを 3 に固定したときの係り受け正解率を、係り受け距離別に示したものである。

この実験では、係り受け距離 1 では、提案方式と南瓜の正解率はほぼ同じとなった。しかし提案方式は、距離が遠くなるに従い、正解率が低下する。1 回のラベリングでは、距離 3 までの係り受けを解析し、文を縮退させるため、距離 4 の係り受けは 2 回目のラベリングで距離 1 として解析される。そのため、係り受け距離 4 の正解率は一時的に上昇している。このグラフは学習／テストともに京都コーパスを用いたものであるが、ブログ+京都コーパスモデルでブログコーパステストセットを解析した場合も同じ傾向であった。

日本語文の場合、多くの係り受けは距離 1 で収まることが多く、ウィンドウサイズを大きくとると、付与すべきラベルの出現比率がバランスしていない。そのため、デフォルト値として距離 1 のラベルが多く付与されることになると考えられる。長距離の係り受けの解析精度向上は今後の課題である。

4 まとめ

本稿では、ブログ等、Web 上で頻繁に現れる準話し言葉を対象に、系列ラベリングによる係り受け解析方法を提案した。提案方式は、書き言葉の係り受け解析器としての精度は南瓜に劣るが、系列ラベリングを用いて柔軟なラベル付けを行うことにより、自分に係る文節も解析可能であることが特徴であるため、ブログ等、フィラーが含まれる文の解析では効果を発揮する。

謝辞

ブログコーパスを作成し、使用させていただいた NTT コミュニケーション科学基礎技術研究所の安田宜仁氏に感謝いたします。

参考文献

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, pages 132–139.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289.

工藤 拓, 松本 裕治. 2002. 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, 43(6):1834–1842.

工藤 拓, 松本 裕治. 2004. 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. 情報処理学会研究報告, 2004-NL-162, pages 205–212.

内元 清貴, 関根 聡, 井佐原 均. 1999. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, 40(9):3397–3407.