4H-05 Cascaded Chunking Model における部分解析済み情報の利用

工藤 拓　　松本 裕治
奈良先端科学技術大学院大学情報科学研究科

1　はじめに

我々は以前にチャンキングの段階適用による高効率な日本語係り受け解析アルゴリズム (以後 Cascaded Chunking Model) を提案した [3]. この手法は, 直後の文節に係るか係らないかという観点のみで決定的に解析を行うため, 文献 [1, 2] といった従来法に比べ, モデル自身が単純で, 実装も容易であり, 高効率である.

本稿では, 部分的に係り受け関係が与えられた状態から全体の係り受け解析が行えるよう Cascaded Chunking Model を拡張することを試みる. さらに, 部分的に付与した係り関係が, 全体の係り受けにどれくらい影響を与えるのか, 簡単な調査を行う.

2　Cascaded Chunking Model

日本語の統計的係り受け解析とは, 非交差条件, 後方参照という二つの制約のもとで, 入力文節列 B に対する係り先パターン列 D の条件付き確率 $P(D|B)$ を最大にする D を求めることと定義できる.

従来法では, それぞれの係り関係は独立であると仮定し, $P(D|B)$ を各二文節間の係り受け確率の積に展開していた. しかし, ある係り関係が他の係り受けに影響を及ぼすこともあり, この独立性の仮定は, 必ずしも適切ではない. さらに, 係り受け確率の推定時や, 実際の解析時に, すべての係り関係の候補を対象としなければならないために, 効率が良いとは言えない.

Cascaded Chunking Model では, 以下の手続きで係り受け解析を行う.

1. 入力文節すべてに対し, 係り受けが未定という意味の O タグを付与する.
2. 文末の文節を除く O タグが付与された文節に対し, 直後の文節に係るか推定する. 係る場合は その文節に D タグに置き換える. 後ろから 2 番目の文節は無条件に D タグ に置き換える.
3. O タグが付与された文節の直後にある文節のうち, D タグが付与されている文節をすべて削除する.
4. 残った文節が一つ（文末の文節）の場合は終了, それ以外は 2. に戻る.

実際の解析例を図 1 に示す. Cascaded Chunking Model は, 直後の文節に係るか係らないかという観点のみで動作し, 必要最低限の係り受け候補のみを推定するために従来法に比べ高効率である. さらに, 文献 [3] では, 係り受け関係そのものを素性として使う**動的素性**を提案し, 係り関係の精度向上に繋がることを示している.

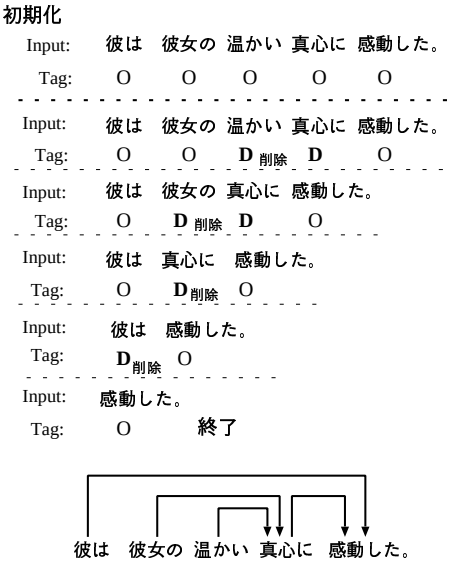

図 1: 係り受け解析例

3　部分解析済み情報の利用

3.1　拡張の目的

部分的に解析済みの情報を利用できるよう Cascaded Chunking Model を拡張する最大の目的は, 「他のシステムとの柔軟な結合」にある. Cascaded Chunking Model は, その解析アルゴリズムの特性上, 遠くの係り関係やトップダウンの情報を利用できないという欠点がある. 例えば, 複文の切り分けや, 並列構造解析などは, 局所的な情報のみを利用するだけでは不十分で, より長いコンテキストを考慮する必要がある. このような場合, Cascaded Chunking Model の弱点を補うことができる他のシステム (具体的には ルールベースシステム, トップダウンパーザ, 人手処理) が高い確信度で出力する係り関係を Cascaded Chunking Model に与えることで, 互いの弱点を補いながら, 全体の係り受け解析精度を向上させることが期待できる.

3.2　拡張手法

従来法では, 個々の係り関係には確率値が付与されるために, 部分的に付与される関係に対し, 十分大きな確

[0]Use of Partial Dependency Structures for Cascaded Chunking Model　Taku Kudoh, Yuji Matsumoto, Graduate School of Infomation Science, Nara Institute of Science and Technology

率値を付与するのみで, 部分情報に対し拡張できる.

その一方で, Cascaded Chunking Model は, 確率値ではなく, 係るか係らないかという単純な二値分類器[1]の結果を利用しながら, 決定的に解析を行う. そのため, 係り関係推定のための二値分類器を書きかえる単純な拡張では追加した情報が考慮されない場合がある. そこで我々は, 係り先候補をあらかじめスタックとして列挙することで, 拡張を行った. 具体的には以下の手続きで処理を行う.

1. 各文節毎にスタックを作成する. スタックには, 部分的に付与された情報を元に, その文節が係る文節の候補を近い順に保持しておく. 部分的に付与されたデータが無い場合は, スタックの中身は, 現在の文節より後方にあるすべての文節となる. 係り先が付与されている場合は, スタックは一つの係り先のみを保持する.

2. スタックを POP して, 空になる場合は無条件に係ける. そうでない場合, 係り受けを推定し, 係る場合には D タグを付与し, スタックを空にする. なお, POP した瞬間に推定せず, スタックトップの文節が直後の文節となった時点で推定する.

3. 文頭あるいは O タグの直後にある D タグがついた文節を削除する.

このアルゴリズムは, 元の Cascaded Chunking Model を完全に包含することに注意されたい.

3.3 部分解析済み情報の追加とその効果

前章のアルゴリズムにより拡張した Cascaded Chunking Model を用い, 部分的に追加した係り関係が他の関係にどの程度影響を与えるか調査を行った.

具体的には, 現実的な設定にするために, 係り関係の推定が困難だと判定された文節を一文につき一つだけ自動的に決定し, その文節の真の係り先を部分解析済みデータとして与えるという設定で実験を行った. 実験に用いたコーパスは京大コーパス (Version 2.0) [4] の一部で, テストには 1 月 9 日分の (1246 文) を用いている. また, 係り関係の困難さは, 以下の方法で算出した.

- 従来法の係り受けアルゴリズムを用い, 上位 N 位までの係り受け解析結果を得る.
- 上位 N 位までの各文節に対し, 係り先までの距離の分散を求める. 分散が一番大きい文節が, 意見の相違が大きく解析困難であると定義する.

また, 一つの文節を無作為に選択し, 部分解析済みデータとして与える手法をベースラインとして実験を行った. 表 1 に実験結果を示す. ただし, 「正解に変更」とは, 追加する前までは不正解だったものが, 正解に変更された個数, 「他に好/悪影響」とは, 追加した文節以外が正解になった個数, 不正解になった個数を示している.

[1]実際には Support Vector Machines を利用している

	分散 ($N = 2$)	分散 ($N = 10$)	無作為
文節正解率 (%)	92.57	92.23	90.38
文正解率 (%)	61.56	57,86	52.08
正解に変更	337	290	116
他に好影響	38	46	11
他に悪影響	6	5	5

表 1: 実験結果

結果から, 少なくとも無作為に選択するよりは, 分散の大きいものを追加する手法が, 精度向上に繋がる事が分かる. また, 係り受けの困難さの判定には, $N = 2$, つまり, 上位 2 位だけを見るだけで充分であった. さらに, 外部から追加する事で, 正解を与えなかった他の係り関係に対しても, 係り先候補を限定する作用が働き, 全体の精度向上に繋がることも分かった.

4 今後の課題

本稿では, 係り先の推定が困難な文節の選択までは自動的に行ったが, 実際の係り先 (答え) は, 教師有りで外部から与えた. 本来ならば, Cascaded Chunking Model の弱点を補うことができるシステムを教師役として使用すべきである. その一方で, 多くの統計的解析器が局所的な情報のみを利用しながら, ボトムアップに解析していく現状を考慮すると, 良い教師役となる解析システムの提案自身が, 今後の課題であると考えている.

5 おわりに

本稿では, 部分的に係り受け解析情報が付与された状況から係り受け解析を行えるよう Cascaded Chunking Model を拡張した. さらに, 冗長解析結果を用い, 係り先の推定が困難と思われる文節を部分解析済みデータとして与える簡単な実験を行った. その結果, 少なくとも無作為に追加するよりは, 精度向上に繋がることが分かった.

参考文献

[1] Taku Kudoh and Yuji Matsumoto. Japanese Dependency Structure Analysis Based on Support Vector Machines. In *Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 18–25, 2000.

[2] 内元清貴, 関根聡, 井佐原均. 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, Vol. 40, No. 9, pp. 3397–3407, 1999.

[3] 工藤拓, 松本裕治. チャンキングの段階適用による係り受け解析. 情報処理学会 自然言語処理研究会 NL142, pp. 9–16, 2001.

[4] 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. 言語処理学会 第 3 回年次大会, pp. 115–118, 1997.