日本語係り受け解析における誤りの類型化と文構造の曖昧性について

山本 悠二 (y_yamamoto@smlab.tutkie.tut.ac.jp) †
増山 繁 (masuyama@tutkie.tut.ac.jp) †

† 豊橋技術科学大学 知識情報工学系

1　はじめに

係り受け解析は，文節間の関係についての基本的な情報を与えるため，自然言語処理における基本技術として認識されている．特に，係り受け解析の正解率向上は，解析後の応用タスクにおける結果に直接影響することが多いため，研究課題として重要な位置を占めている．しかしながら，解析器の係り受け正解率は新聞記事を対象とした場合，現状で約 91%程度で飽和状態となっている．このように解析精度が頭打ちになっている理由が，係り受け解析器のルールや素性，解析戦略に改善の余地があるためなのか，日本語文に内在する依存構造の曖昧性によるものであるのかは，今のところ，はっきりとしていない．そのため，係り先が一意に決定できるか否かについての検証を含めた，係り受け誤りの原因を網羅的に解析することが重要になる．

これまで，特に，統計的係り受け解析における誤り解析では，あらかじめ定められた素性集合から素性を抜き取って正解率を調査するものが広く用いられてきた．しかし，これは素性が有効に働くかどうかを調べるものであって，どのような素性をさらに追加すれば解析精度がよくなるかについては，この方法で調べることはできない．また，論文 [1] では，2 つの解析器間における係り元品詞 (活用形を含める) を対象とした係り受け正解数の比較を行っている．これは二者の解析器の品詞ごとの得意/不得意を評価するのには適しているが，単体の解析器について何が不得意であるかを直接把握することは難しい．類似の方法として，単体の解析器について係り元品詞 (活用形を含める) を対象として誤りの個数を調べることが考えられるが，品詞の異なり数が多いため，類型化して評価する必要がある．

また，人手による詳細な誤り解析については，研究 [4] があるが，評価に時間がかかるため対象とする文数を限定しなければならないところに問題がある．実際に，先の研究では対象としている文数が数百に留まっている．そのため，係り受け解析の誤りを網羅的に解析するためには，ある程度，自動的に誤りの分類を行う必要がある．

本論文では，統計的日本語係り受け解析誤りの原因を網羅的に検証することを目的として，係り受け解析の誤りを類型化した．そして，誤りがどの類型に属するかについての自動分類のルールを作成し，実際の係り受け解析器に適用した際のルールの網羅率，ならびに，類型ごとの誤りの分布を調べた．さらに，それぞれの類型について人手での評価を行い，実際に係り先が一意に決定できるか否かについて検証を行った．

2　日本語係り受け解析における誤りの類型化

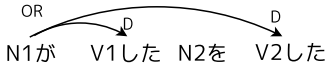
文献 [5] (p.186) では，日本語文の依存構造における曖昧性を 4 つに類型化している．本論文では先の類型化を拡張することで，日本語係り受けの誤りを類型化した．相違点は以下の 5 点である．

1. 係り受け解析において，係り関係 (従属/並列など) の同定は，係り先の同定を行った後の処理であると設定する場合が多い．そのため，係り受け解析器は係り先の同定のみ行うことを想定する．係り関係についてはアノテーションデータを参照して分類する．
2. 格要素の係り先の曖昧性に加えて，主に提題表現として用いられる助詞「は」「も」の係り先の曖昧性を加えた．
3. 並列構造の範囲の曖昧性が格要素の係り先により生じるものか，もしくは，並列節により生じるものかに分けた．
4. 事前実験で並列構造における助詞「は」「も」の係り先の誤りは，解析器が部分構造を特定することができなかったものが 7 割ほど占めていたため,『並列構造における助詞「は」「も」の部分構造の特定の誤り』に限定した．
5. 事前実験で，人手で誤りの類型化を行ったところ，接続詞，副詞の係り先の曖昧性が見られたため，これらを加えた．

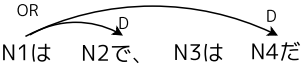
類型化したものと，それぞれの文例を以下に示す．

1. 連体修飾の係り先の曖昧性

2. 従属関係での格要素[1]の係り先の曖昧性



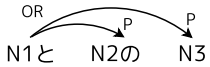
3. 従属関係での助詞「は」「も」の係り先の曖昧性



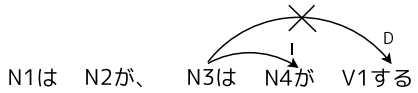
4. 従属節の係り先の曖昧性

[1] 時間表現 [3] も含む．

5. 並列関係での格要素の係り先の曖昧性



6. 並列構造における助詞「は」「も」の部分構造の特定の誤り

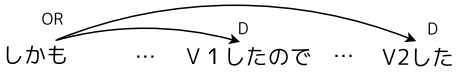


7. 並列節の係り先の曖昧性[2]

8. 同列の並列構造における係り先の曖昧性

9. 副詞の係り先の曖昧性[3]

10. 接続詞の係り先の曖昧性



3 ルールに基づく誤りの自動分類

1節でも述べたように，実際の係り受け解析の誤り解析では，誤りの箇所をできるだけ網羅できることが望ましい．本論文では，論文 [6] をベースにして誤りの類型を自動分類するルールを作成した．

まず，自動分類を行うための前提について述べる．データとして，あるコーパスに対して，係り先・係り関係のアノテーションを人手で付与したものと，係り受け解析器によって係り先を同定したものが与えられているとする．そして，人手で付与した係り先と，解析器によって付与した係り先が異なるものを対象として，誤りの類型を自動分類する．具体的には，それぞれの解析誤りについて，ルールで定めた係り元語形，係り先主辞，係り関係の属性値が合致する場合[4]に，判別された誤りの類型を出力する．ただし，ここでの語形とは，文節内の品詞が特殊となるものを除き，文末に一番近い形態素のことを，また，主辞とは文節内で品詞が特殊，ないし，助詞となるものを除いた文末に一番近い形態素のことを指す．

以下に誤りの類型を自動分類するルールを示す．

1. 連体修飾の係り先の曖昧性 (係り関係: 従属)
   係り元語形 : 連体詞，助詞「の」，活用形が基本形/タ形
   係り先主辞 : 名詞，接尾辞で名詞性のもの
2. 従属関係での格要素の係り先の曖昧性 (係り関係: 従属)
   係り元語形 : 助詞・格助詞，名詞・時相名詞，接尾辞「年」「月」「日」「時」
3. 従属関係での助詞「は」「も」の係り先の曖昧性 (係り関係: 従属)
   係り元語形 : 助詞「は」「も」
4. 従属節の係り先の曖昧性 (係り関係: 従属)
   係り元語形 : 助詞・接続助詞，活用形が連用形/基本条件形

   もしくは，

   係り元語形 : 名詞・サ変名詞
   係り先主辞 : 動詞
5. 並列関係での格要素の係り先の曖昧性 (係り関係: 並列/部分並列)
   係り元語形 : 助詞・格助詞，名詞・時相名詞，接尾辞「年」「月」「日」「時」
6. 並列構造における助詞「は」「も」の部分構造の特定の誤り (係り関係: 部分並列)
   係り元語形 : 助詞「は」「も」
7. 並列節の係り先の曖昧性 (係り関係: 並列)
   係り元語形 : 活用形が連用形，助詞・接続助詞，活用形が基本条件形

   もしくは，

   係り元語形 : 名詞・サ変名詞
   係り先主辞 : 動詞
8. 同列の並列構造における係り先の曖昧性 (係り関係: 同列)
   係り元語形 : 助詞・格助詞，接尾辞「年」「月」「日」「時」

   もしくは，

   係り元語形 : 接尾辞で名詞性のもの，助詞「など」，名詞
   係り先主辞 : 名詞，接尾辞で名詞性のもの
9. 副詞の係り先の曖昧性 (係り関係: 従属)
   係り元語形 : 副詞，名詞・副詞的名詞
10. 接続詞の係り先の曖昧性 (係り関係: 従属)
    係り元語形 : 接続詞

[2] アノテーション・マニュアル [2] によると，文中に 3 つ以上の並列部分がある場合,「各部分の主辞からその右隣の並列部分の主辞にリンクをはる」こととなっている．つまり，文例では一般的には V2 が係り先であると定めることが妥当である．しかし，語彙的に見て V2 よりも V3 を係り先として定める場合がある．

[3] 副詞的名詞も含める．

[4] ただし，係り先主辞については，人手で付与された係り先にある主辞と，解析器によって付与された係り先にある主辞の両方の属性値が合致する必要がある．

4 実験

実際の係り受け解析器に自動分類を適用し，ルールの網羅率，ならびに，類型ごとの誤りの分布を調査した．

4.1 実験設定

使用する係り受け解析器は，後方文脈，ならびに，係り先候補間の相対的な距離を考慮できる点から論文 [7] で示されているものを使用した[5]．京都テキストコーパス 4.0 [6]を以下の 3 つに分けて実験を行った．

- 訓練データ: 一般記事[7] 1 月 1，3-11 日，社説 1-8 月，合計 24,280 文，234,639 文節
- 開発データ: 一般記事 1 月 12，13 日，社説 9 月，合計 4,833 文，47,571 文節
- 評価データ: 一般記事 1 月 14-17 日，社説 10-12 月，合計 9,284 文，89,874 文節

機械学習を行う際に使用した素性は，CaboCha 0.53 [8]中にある素性抽出プログラム selector.pl の出力によるものである．また，他の語彙的な素性の追加は固有のコーパスに過度に依存する可能性があるため，使用していない．使用したカーネルは 3 次の多項式で，学習器のコストパラメータを 1 と定めた．また，反復回数の設定は，1 から 10 回までの反復回数のうちで開発データの係り受け正解率が最も高くなった値である 6 回目を使用した．

評価データによる係り受け正解率は 91.24 % (73529 / 80590)，文正解率は 55.14 % (5119 / 9284) であった[9]．なお，ここでの係り受け正解率は，文末の 1 文節を除くすべての文節に対して正しく係り先が同定できたものの割合，文正解率は，文単位で全体の文節の係り先が正しく同定できたものの割合を示す．

4.2 実験結果

類型ごとの誤り数を表 1 に示す．結果から全体の係り受け誤りの 8 割程度がルールにより網羅できることが確認できる．

もう一つの実験として，各類型から誤りを無作為に 30 個取り出し，正しく自動分類できているか否かについて人手で評価を行った．結果を表 2 に示す．結果から「該当せず」を除けば，各類型は 8 割以上の正解率であった．他の類型と比べて正解率が低い類型 4 と 5 の誤り原因は，類型 4 では複合辞を抽出したこと (全体の 50%)，類型 7 では並列句を誤って抽出したこと (全体の 80%) であった．また,「該当せず」に誤って分類された主な類型は，類型 2(全体の 50%)，類型 8(全体の約 29%) であった．誤分類についてはルールを追加することで正しく分類できると考える．以上のことから全体として，類型の自動分類の再現率には改善の余地があるものの，精度は十分高いと考える．

表 1 を再度確認すると，誤り数が最も多い類型は「従属関係での格要素の係り先の曖昧性」である．この理由は，統計的係り受け解析では，格要素がどの述語に係るかを限られた学習データから推定するため，語彙による係り受けの選好を考慮することが難しいためであると考えられる．そのため，格フレーム辞書などの外部知識の併用が必要であると考える．

表 1　類型ごとの誤り数 (自動)

誤り類型	個数
1	818
2	1674
3	1273
4	646
5	263
6	27
7	650
8	46
9	444
10	53
該当せず	1167

表 2　自動分類結果における評価 (人手)

誤り類型	正しく分類	誤って分類
1	30	0
2	30	0
3	29	1
4	24	6
5	30	0
6	27	0
7	24	6
8	30	0
9	30	0
10	30	0
該当せず	16	14

4.3 誤り類型ごとの係り先の評価

誤り類型ごとに，人手による係り先のアノテーションが文内で一意に決定できるかについて調べた．4.2 節で示した「人手での評価」で使用した誤り (各類型から誤りを無作為に 30 個取り出したもの) のそれぞ

[5] 同様のアイデアに基づく係り受け解析モデルは岩立らも独立に提案している [8].
[6] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
[7] ただし, 以下の ID を持つ文は文節番号とその文節の係り先番号が同一であるというタグ付けの誤りがあったため, 訓練データから除外した; 950101159-010, 950106177-017, 950106192-002.
[8] http://chasen.org/~taku/software/cabocha/
[9] 同様のデータセットについて CaboCha 0.53 (3 次の多項式カーネル，コスト 1 で学習) で評価したところ，係り受け正解率 90.27 % (72747 / 80590)，文正解率 52.60 % (4883 / 9284) であった．

表 3　係り先の正しさの評価 (人手)

誤り類型	アノテーションが正しい	解析器が正しい	両方間違い	一意に決まらない
1	21	5	2	2
2	18	6	3	3
3	20	3	2	5
4	23	3	1	3
5	25	3	1	1
6	27	0	0	0
7	25	1	3	1
8	22	5	2	1
9	18	7	0	5
10	18	6	2	4
該当せず	21	6	1	2

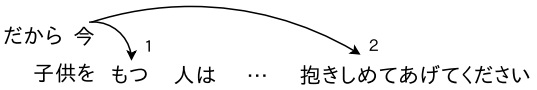


図 1　係り先が一意に決まらない例

れの係り先について「アノテーションが正しい」,「解析器が正しい」,「両方間違い」,「一意に決まらない」の 4 つに人手で分類した．結果を表 3 に示す．全体の誤りのうちで一意に決まらないものが 1 割以上含まれている類型は，類型 2，類型 3，類型 4，類型 8，類型 9 であった．類型 2 における例を図 1 に示す．矢印の先に数値 1 を付けているものがアノテーションの係り先，数値 2 を付けているものが解析器の係り先をそれぞれ表す．この文は時間表現が後続する 2 つの述語のどれに係るのかが一意に定まらない．同様の問題は，副詞，接続詞について多く見られた．以上のことから，実用上，確実な係り受けを必要とするのであれば，副詞と接続詞については，係り受けのアノテーションとして具体的な係り先を定めずに，後続する述語のどれかであるというタグを付与する程度に留めるといった方向性があることが示唆される．

5　まとめと今後の課題

本論文では，統計的日本語係り受け解析の誤りの原因を網羅的に検証するために，係り受け解析の誤りの類型化，ならびに，類型の自動分類のルールを作成した．そして，実際の係り受け解析器で網羅率，ならびに，類型ごとの誤りの分布を調べた．さらに，それぞれの類型について人手での評価を行い，実際に係り先が一意に決定できるのかについて検証を行った．今回の人手での評価に用いた誤り数は，各類型でそれぞれ 30 個と少数ではあるものの，統計的日本語係り受け解析の誤りのうちで係り先が一意に定まらないことによるものが含まれることが確認された．

今後の課題として，人手での評価に用いる誤り数を増やし，係り先の曖昧性に起因する解析誤りについての詳細な調査を行いたい．また，誤り類型の自動分類における再現率向上，および，係り先の曖昧性に対応したアノテーション・ルールの考案についても検討していきたい．

謝辞

本研究は文部科学省グローバル COE プログラム「インテリジェント　センシングのフロンティア」の援助により行われた．

参考文献

[1] 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.

[2] 河原 大輔，笹野 遼平，黒橋 禎夫，橋田 浩一，“格・省略・共参照タグ付けの基準”

[3] 黒橋 禎夫，居蔵 由衣子，坂口 晶子，“形態素・構文タグ付きコーパス作成の作業基準”

[4] 河原 大輔，“日本語係り受け解析の誤り分析と精度上限に関する考察，” 第 3 回 NLP 若手の会，2008.

[5] 長尾 眞，佐藤 理史，黒橋 禎夫，角田達彦．“自然言語処理 (岩波講座 ソフトウェア科学 15)”．岩波書店，1996.

[6] 黒橋 禎夫，長尾 眞，“並列構造の検出に基づく長い日本語文の構文解析，” 自然言語処理，vol.1，no.1，pp.35-57，1994.

[7] 山本 悠二，増山 繁，“係り先候補の相対的な距離を反映した統計的日本語係り受け解析，” 情報処理学会研究報告 自然言語処理，no.113，pp.15-22，2007.

[8] 岩立将和, 浅原正幸, 松本裕治，“トーナメントモデルをを用いた日本語係り受け解析，” 自然言語処理，vol.15，no.6，pp.169-185，2008.