基般

3.2 構文解析

河原 大輔（京都大学）

構文解析とは

文は語の並びであり，語と語の間には構造がある．語間の構造は**構文**と呼ばれ，これを解析する処理は**構文解析**と呼ばれる．構文の表現形式の代表例として，**句構造**と**係り受け構造**がある．これまで，英語では句構造を中心に，日本語では係り受け構造を中心に研究が進められてきた．本稿では，日本語の係り受け構造の解析，すなわち**係り受け解析**について述べる．

係り受け構造は，2つの要素間（単語や句）の係り受け関係の集合として文をとらえたものである．日本語における係り受け関係は，通常，文節を単位とし，2つの文節間の関係と考える．

この文は，「女の子が」「クロールで」「泳ぐ」という3つの文節からなり，係り受け構造は，「女の子が → 泳ぐ」「クロールで → 泳ぐ」という2つの係り受け関係からなる．このように，日本語では，文末の文節（上記の例では「泳ぐ」）を除き，係り元は右側にただ1つの係り先を持つ．つまり，日本語の係り受け関係は左から右への一方向である．また，日本語では原則として，各係り受け関係は互いに交差しないという特徴を持つ．

係り受け解析は，入力文に対してこのような係り受け構造を明らかにする処理である．係り受け解析が難しい例を次に示す．

(2) クロールで 泳いでいる 女の子を 見た
(3) 望遠鏡で 泳いでいる 女の子を 見た

この2文では，「クロールで」と「望遠鏡で」の係り先が異なっており，(2) では「見た」，(3) では「泳いでいる」に係る．これらを正しく解析するためには，後述する語彙的選好知識が必要である．

解析手法

係り受け解析の手法としては，以前は係り受け規則，選好を人手で記述していたが，現在では機械学習をすることが主流である．機械学習は，正解係り受け情報を人手で付与したタグ付きコーパスを用いて行う．日本語の係り受け構造のタグ付きコーパスとしては，京都大学テキストコーパス☆1がよく利用されている．

機械学習手法は，このようなコーパスから，「名詞は動詞に係ることができる」「距離が近い文節に係りやすい」といった文法的な傾向を学習している．しかし，タグ付きコーパスは数万文規模であり，「クロールで → 泳ぐ」や「望遠鏡で → 見る」のような語彙的な選好を学習することは難しい．この問題については後述する．

日本語係り受け解析のツールとしては，KNP☆2, CaboCha☆3, J.DepP☆4などが公開されている．解析精度はいずれも90%程度である．

何が難しいのか

係り受け解析の10%の誤りのうち，次の3つが主な誤りである．

1. 並列構造の誤り

並列構造は，日本語文中に頻出する表現であり，

☆1 http://nlp.ist.i.kyoto-u.ac.jp/index.php? 京都大学テキストコーパス
☆2 http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
☆3 https://taku910.github.io/cabocha/
☆4 http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

係り受け構造に直結しているため，並列構造の認識およびスコープ同定を誤ると，係り受け解析も大きく誤ることになる[1]．次の例では,「性格と,」と「真空調理を」が並列と解析され誤っているが,「性格と,」は,「〜と」をとりやすい「かみ合っての」に係るのが正解であり，並列構造を構成していない．

(4) まさに，仕出しという仕事の**性格と，**真空調理を× 使ってできる合理化，能率化がうまくかみ合っての○ 成功である．

この文における並列構造は,「合理化」と「能率化」の部分のみである．なお，本稿中の例文の表記として，太字部を係り元，○下線部を正解の係り先,×下線部を自動解析結果の誤り係り先とする．

2. 正解タグ付けの問題

係り受け構造の正解タグ付けは，仕様として，文節ごとに1つの係り先を決めている．ある文節が，意味的には複数の文節と関係があり，それぞれに係ることができても，正解としてはタグ付け基準に従って1つに決める必要がある．そのため,自動解析の係り先が意味的には誤りではなくても,正解係り先とずれることがある．また，実際問題として正解タグ付け時の誤りも存在する．次の例では,「期間中は」が○下線部と×下線部のどちらとも意味的に関係があるが，正解と自動解析で係り先が異なっている．

(5) この**期間中は**交通規制などで皆さんにご不便をおかけしますが，○ ご協力をお願いします．×

3. 語彙的選好知識のカバレッジ不足

語彙的な選好知識がないと，正しい係り先を解析することが難しい場合がある．次の例では「旅行などで,」の係り先として,「参詣する」が他の係り先候補よりも語彙的に優勢であると考えられる．

(6) **旅行などで，**日頃馴染みのない寺院や神社に参詣する○ 際にも，幅広く寺社札を受けて来たと× いうことなのでしょう．

この問題は，述語がとり得る名詞の知識（格フレームと呼ばれる）を大規模コーパスから学習することによってある程度解決できる[2]．

今後の展開

構文解析は，意味解析や言語処理アプリケーションの多くが利用する基盤技術であるため，さらなる精度向上が望まれる．上記の「正解タグ付けの問題」があるため，100% とはいかないまでも，95% 程度は目標として設定できると考えられる．今後，この目標を達成するために，大規模な知識の獲得および利用や，ニューラルネットワークによる学習などが盛んに行われると思われる．また，さまざまなドメインのテキストに対して頑健に解析できることも求められており，このためにも，対象ドメインのテキストからの知識獲得が重要となる．

参考文献

1) Kurohashi, S. and Nagao, M. : A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. Computational Linguistics, 20(4) : pp.507-534 (1994).
2) Sasano, R., Kawahara, D. and Kurohashi, S. : The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis, In Proceedings of NAACL-HLT 2009, pp.521-529 (2009).

(2015年11月2日受付)

河原 大輔（正会員）dk@i.kyoto-u.ac.jp
2002年京都大学大学院情報学研究科博士課程単位取得認定退学．東京大学大学院情報理工学系研究科学術研究支援員，(独）情報通信研究機構主任研究員を経て，2010年より京都大学大学院情報学研究科准教授．自然言語処理，知識処理の研究に従事．博士（情報学）．