CCG パーザを用いた未知語の統語範疇自動推定

田中リベカ　（指導教員：戸次大介）

1　序論

自然言語処理においてはその興味の対象が深い意味に移行してきているが、人間による言語の意味理解の仕組みを探究するにあたって重要な役割を担うタスクの１つに、統語解析が挙げられる。統語解析を形式文法に基づいて行う上で課題となるのは実テキスト中に表れる未知語への柔軟な対応である。組合せ範疇文法 (CCG:[1]) などの語彙化文法においては辞書と規則から文が生成されるため、辞書に登録されていない未知語に対応することが、統語解析を自動化する鍵となる。

範疇文法では未知語の範疇に対して強力な推論がはたらくため、Watkinson ら [2] や Yao ら [3] により、未知語の統語範疇の自動推定が試みられている。しかしそれらのアルゴリズムでは未知語の出現数が 1 文当たり 1 個のみであることが条件となっており、文中に出現する未知語の個数が予測できない実テキストに対応するには課題が残ると考えられる。また、あらかじめ未知語に割り当てる統語範疇のラベルの集合を具体的に決定しておき、それらの統語範疇ラベルから確率の高いものを選択するという手法 [4] もとられている。しかしこの方法では、ラベルとして用意していなかった統語範疇には対応できないという問題がある。

本研究では、CCG の文法理論に基づき、文中の任意個の未知語に対しパーザを用いた統語範疇推定を行う。

2　組合せ範疇文法（CCG)

CCG においては、基本的な統語範疇として n（名詞)、np（名詞句)、s（文）等が用意されている。それ以外の複合的な統語範疇は、これらの基本的な統語範疇と演算子「/」「\」によって「s/np」「$s\backslash np$」のように表される。CCG には統語範疇を組み合わせるために以下をはじめとする 9 つの規則が存在する：

関数適用規則	$(>)$	$X/Y\ Y$	$\Rightarrow$	X
	$(<)$	$Y\ X\backslash Y$	$\Rightarrow$	X
関数合成規則	$(>B)$	$X/Y\ Y/Z$	$\Rightarrow$	X/Z
	$(<B)$	$Y\backslash Z\ X\backslash Y$	$\Rightarrow$	$X\backslash Z$

これにより、“John runs shops” という文の導出は以下のように証明木で表される：

$$\dfrac{\dfrac{\textbf{John}}{np}\quad\dfrac{\dfrac{\textbf{runs}}{(s\backslash np)/np}\quad\dfrac{\textbf{shops}}{np}}{s\backslash np}(>)}{s}(<)$$

CCG の規則は少数であるため、実質的には辞書の獲得が文法の獲得に相当する。すなわち、CCG で文の導出を行うには語の統語範疇が既知であることが必要である。ここで、各語の統語範疇は以下のような性質を満たすものでなければならない：

- 各語の統語範疇はその周辺の語といずれかの規則を用いて統合される（ただし一部例外もある）
- 証明木の根は s（文）となる

この制約により、CCG では統語範疇に対して強力な推論が働き、周囲の語から情報を得ることが可能である。

3　手法

処理の概要を図 1 に示す。処理工程はコーパスと CCG の規則・辞書、パーザ、解決定器からなる。コーパスは英文からなり、文中に未知語（辞書に定義されていない語）が出現する文を含む。CCG 規則としては、現時点では先述した関数適用規則と関数合成規則のみを用いている。パーザはコーパスの文を 1 文ずつ入力として受け取り、各文の構文解析を行った結果として未知語の統語範疇の候補を出力する。解決定器はこの候補から解を絞り込む処理を行う。また、これらの実装は論理型言語 LiLFeS で行った。

3.1　アルゴリズム

未知語推定処理は (1) 候補カテゴリの全探索 (2) 解の決定の 2 段階で行う。

(1) 候補カテゴリの全探索

未知語の統語範疇としてありうる全ての候補を、CKY 法を採用したパーザ [5] を用いて探索する。

通常 “John runs fast.” のような未知語を含まない文の構文解析においては、CKY 法では文中の各語について辞書に登録されている統語範疇を参照し、ボトムアップに文構造の全可能性を探索する。この際、ある語に対して辞書に複数の統語範疇が登録されていた場合は、複数の可能性全てについて導出を試みる。

本手法では、“John runs X.” のような辞書に登録されていない未知語 “X” を含んだ文について、CKY 法による統語解析を行う。このとき、未知語の統語範疇は変数 x とする。これにより、文 “John runs X.” のとりうる文構造を全て探索する中で、逆に “John runs X.” が文となる場合の変数 x の値の候補、つまり未知語 X の統語範疇の候補を全て調べることができる。

(2) 解の決定

上のように “John runs X.” という特定の文における未知語 X の統語範疇を調べるだけでは、X が “John runs fast.” のような副詞なのか、または “John runs shops.” のような名詞なのかは判断できない。未知語の統語範疇の候補を絞り解を決定するには、その語が他の文でどのように振舞うのかを考慮する必要がある。

本手法では、コーパス上の全ての文について「その文における未知語の統語範疇の候補集合」を取得した後、表層的に同じ語については、これらの候補集合を文間で比較する。これにより、複数の文に共通する未知語 X の統語範疇の候補が取得できる。言い換えると、未知語 X を含むコーパス上の文の全てが統語解析可能となるような統語範疇 x の形がわかるのである。

この段階でなお複数の候補が残る場合、それに含まれる基本的な統語範疇の数が最も小さいものを解として採用する。この段階での解候補はどれも未知語 X の統語範疇として相応しいといえるが、人間は望ましい解候補に絞り込むために何らかの基準を持っていると考えられる。現時点では統語範疇の長さをその基準の代わりに用いているが、将来的にはより本質的な基準を用いるよう改良を加える方針である。

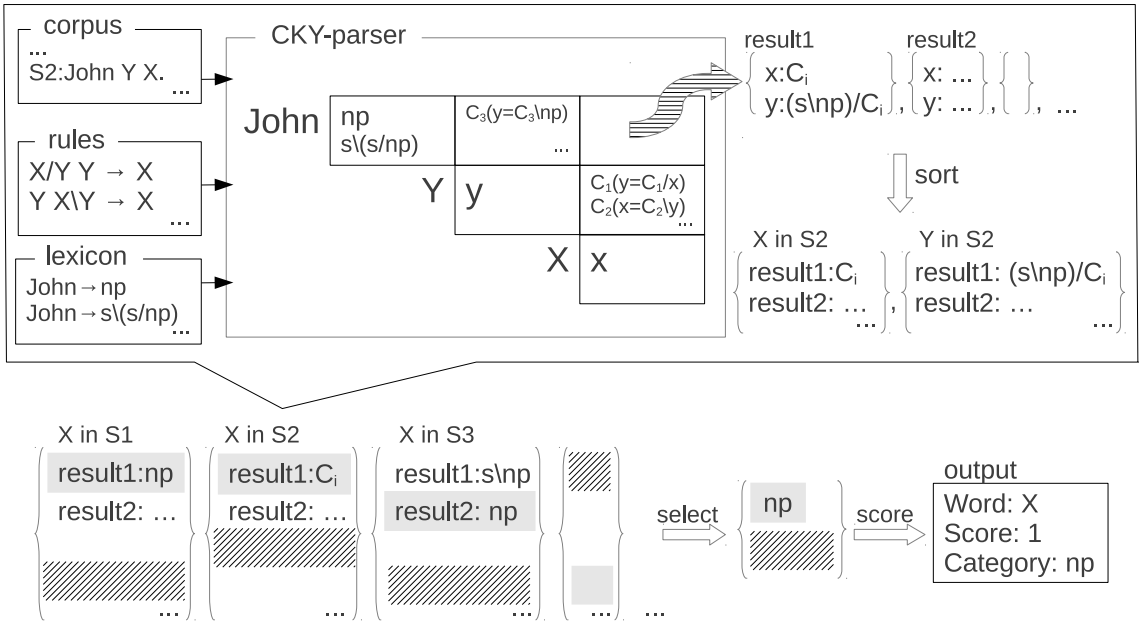

図 1: 未知語推定の処理工程

3.2 複数個の未知語の出現

先述した手法を用いて、文中に複数個の未知語が出現する場合にも統語範疇を推定することが可能である。

(1) 候補カテゴリの全探索について

“John Y X.” のように文中に未知語 X、Y が出現する場合、X の統語範疇は Y の統語範疇に依存し、またその逆も成立する。そのため、候補として取得された統語範疇それ自体に変数が出現しうる。統語範疇に変数が含まれるということは原理的に統語範疇を決定することができないということを意味しており、“John Y X.” という文から未知語 X、Y の統語範疇として多数の可能性が挙げられることを考えると、妥当な現象であると言える。

(2) 解の決定について

“John Y X.” における未知語 X のように統語範疇が変数を含むものであっても、他の文における統語範疇の候補と比較することで、可能性を絞ることができる。たとえば、未知語 X の統語範疇の候補が変数 $*$ を含み、「$*/np$」という形であるとする。このとき、他の文において候補「s/np」が有力であったとすると、$* = np$ と解釈するのが都合が良いと言える。また、他の文において「s/s」が候補に挙げられていた場合、「$*/np$」とは異なるカテゴリであると判断できる。このように本手法では、変数の単一化を用いることで統語範疇の比較を行い、複数の文に共通する統語範疇の候補を取得する。なお複数の候補が残る場合は先述したスコア付けを行うが、その際、変数にはその特性に応じて 1 または 2 の長さを与える方針をとる。

4 今後の課題

現在は、複数の文に共通する統語範疇の候補を取得する際、共通する範疇が少なくとも 1 つは存在することを仮定している。しかし、1 つの動詞が自動詞と他動詞の用法をもつなど、複数の統語範疇が辞書に登録されるべき場合がある。このような場合には、全ての文に共通する統語範疇の候補はなく、未知語の統語範疇が「自動詞、または他動詞」という複数の可能性からなるときのみコーパス上の全文が生成可能となる。今後、このような複数用法の統語範疇取得も可能にする必要がある。

これに関連して、既に辞書に登録されている語の新用法が発見された場合についても、辞書に追加するための機構が必要であると考えられる。

5 結論

本研究では、文中の任意個の未知語に対し CCG パーザを用いた統語範疇推定を行った。今後、さらに精度の高い推定を可能にするとともに、大規模データに対して本手法を適用し評価を行う方針である。

参考文献

[1] Steedman Mark J. *Surface structure and interpretation*, Vol. 30. The MIT press, Cambridge, MA, 1996.

[2] S. Watkinson and S. Man. Unsupervised lexical learning with categorial grammars. In *ACL'99: Workshop in Unsupervised Learning in Natural Language Processing*, 1999.

[3] Xuchen Yao, Jianqiang Ma, and Sergio Duarte. Unsupervised syntax learning with categorial grammars using inference rules. In *Proceedings of the 14th Student Session of the European Summer School for Logic, Language, and Information*, 2009.

[4] Miyao Yusuke. *From Linguistic Theory to Syntactic Analysis: Corpus-Oriented Grammar Development and Feature Forest Model*. Ph.D thesis, Department of Computer Science, The University of Tokyo, 2006.

[5] 尾崎博子. 範疇文法と部分方向性組み合わせ論理の Curry-Howard 対応に基づく統語解析. 修士論文, お茶の水女子大学, 2013.