4B1-6

CCGパーザを用いた未知語の統語範疇自動推定

Inference of Syntactic Categories by Using a CCG parser

田中リベカ[*1] 宮尾祐介[*2] 戸次大介[*1*2]
Ribeka Tanaka Yusuke Miyao Daisuke Bekki

[*1]お茶の水女子大学大学院人間文化創成科学研究科
Graduate School of Humanities and Sciences, Advanced Science, Ochanomizu University

[*2]国立情報学研究所
National Insutitute of Informatics

For parsing with formal grammars, detection of unknown words is an important task. In Combinatory Categorial Grammar (CCG), a syntactic category of a word can be inferred from other words around it by using inference rules. However, such a bottom-up analysis yields too many potential candidates and it makes us difficult to determine the appropriate category. Even if linguists can determine a syntactic category for a word, it's not clear why they choose it. From both engineering and linguistic points of view, it is important to reveal factors that determine the appropriate syntactic categories. In this paper, we propose one of those factors and report the results of our experimental validation using a CCG parser.

1. はじめに

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始めているが、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現れる未知語への柔軟な対応である。組合せ範疇文法 (CCG:[Steedman 96]) 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周囲の語の統語範疇と推論規則から強力な推論が働くため、文中に現れる未知語の統語範疇を逆算的に求めることが理論上は可能である。このような性質を生かして周囲の語との関係から未知語の統語範疇を自動推定する試みがなされているが [Yao 09]、現在はあらかじめ未知語に割り当てる統語範疇のラベル群を具体的に用意しておく supertagging の手法で未知語に対処するアプローチが主流になっている [Ninomiya 06, Clark 07]。

このような手法がとられている背景には、範疇文法の性質を利用した逆算的な推論だけでは未知語の統語範疇の候補を絞り込むことができないという事情がある。一般に言語学者がある言語に統語範疇の集合を与える際には、範疇文法で記述できる全ての可能性を使用しているわけではなく、自らが思い描く言語観に沿った振る舞いをするよう、暗黙のポリシーに基づき統語範疇を決定する。これに対し逆算的に統語範疇を推論する場合には、普段は言語学者が無意識に候補から除外するような統語範疇も含めて全ての可能性が計算されるため、言語学者のポリシーが明示化されていなければ、それ以上解候補を絞ることは困難となってしまうのである。

supertagging の手法においては、言語学者の暗黙のポリシーに基づいて作成された統語範疇のラベル群をあらかじめ用意しておくことで、逆算的な推論による候補の増大を回避している。しかし将来的に複合表現や方言などの語についての深い分析に踏み込んでいく上では、既存の統語範疇とは異なる未知の統語範疇が出現することが予想され、予測不可能な統語範疇にも対処できることが望ましい。

そこで本研究では、多数の候補の中から適切な統語範疇へと絞り込むための制約、つまり言語学者が語に統語範疇を与える際のヒューリスティクスを顕在化することで未知語の統語範疇の自動推定を可能にすることを目標とする。本稿では、未知語を含む簡単な英語の文について CCG パーザを用いて未知語の統語範疇の候補を具体的に求め、その後制約を 1 つ仮定し、候補範疇の候補を絞り込むことを試みた。

連絡先: 田中リベカ, お茶の水女子大学大学院人間文化創成科学研究科理学専攻戸次研究室, 東京都文京区大塚 2-1-1, tanaka.ribeka@is.ocha.ac.jp

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含む 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)	$Y\backslash Z\ X\backslash Y$	$\Rightarrow$	$X\backslash Z$
型繰上げ規則	(>T)	X	$\Rightarrow$	$T/(T\backslash X)$
	(<T)	X	$\Rightarrow$	$T\backslash(T/X)$

なお、本研究では型繰上げ規則を規則としては採用しておらず[*1]、代わりに辞書に初めから型繰上げ規則適用後の統語範疇を登録することで対応している。

これにより、“John sees Vincent” という文の導出は以下のような証明木で表される：

$$\dfrac{\dfrac{\textbf{John}\ T/(T\backslash NP) \qquad \dfrac{\textbf{sees}\ (S\backslash NP)/NP \qquad \textbf{Vincent}\ T\backslash(T/NP)}{S\backslash NP}\,(<)}{S}\,(>)}$$

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

*1 型繰上げ規則は理論上どのような統語範疇に対しても繰り返し適用することが可能であり、推論規則として採用するとナイーブな実装ではパーザの処理が停止しないという問題があるためである。

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：

1. 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）
2. 証明木の根は S（文）となる

これらの制約により、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。上記制約を満たすものが CCG における語の統語範疇の候補になりうるといえる。

3. 逆算的な推論による解候補の探索

先述した CCG の 2 つの制約を満たす統語範疇全体を解候補として求めるため、本研究では、入力文、CCG の規則（関数適用規則、関数合成規則、等位接続規則の 5 規則のみ）・辞書、パーザから構成される解析器を用いる。解析器の実装は、論理型言語 LiLFeS[Makino 97] で行った。

まず、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 [尾崎 13] で列挙する。通常統語範疇が全て既知である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 X Vincent.” のような、統語範疇が未知で辞書への登録がない未知語 “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 であると仮定して構文解析を行う。ここでは、“John” と “Vincent” については辞書に登録されている統語範疇 $T\backslash(T/NP)$、$T/(T\backslash NP)$（ただし T は変数）が使用され、辞書に登録されていない未知語 “X” の統語範疇には変数 x を使用することになる。この初期値を用いて CKY 法で構文解析を行うと、変数 T や x は解析の過程で単一化を繰り返す。導出結果の例を以下に示す。

導出例 1.

$$\dfrac{\mathbf{John}\ T_1\backslash(T_1/NP) \qquad \dfrac{\mathbf{X}\ (T_2\backslash T_1)/NP \qquad \mathbf{Vincent}\ (T_2\backslash T_1)\backslash((T_2\backslash T_1)/NP)}{T_2\backslash T_1}(<)}{T_2\backslash T_1/NP}(<B)$$

導出例 2.

$$\dfrac{\mathbf{John}\ T_3/(T_3\backslash NP) \qquad \dfrac{\mathbf{X}\ (T_3\backslash NP)/NP \qquad \mathbf{Vincent}\ (T_3\backslash NP)\backslash((T_3\backslash NP)/NP)}{T_3\backslash NP}(<)}{T_3}(>)$$

入力は文であるため、構文解析結果の統語範疇と文の型 S との単一化が成功した場合のみ、その時の変数 x の値を結果として取得する。上記の例では $T_2\backslash(T_1/NP)$ は S と単一化できない。一方、T_3 は S と単一化可能で $T_3 = S$ となるため、その時の変数 x の値に該当する $(T_3\backslash NP)/NP = (S\backslash NP)/NP$ を未知語 X の統語範疇の候補とする。このプロセスにより、文 “John X Vincent.” のとりうる文構造を全て探索する中で、逆に “John X Vincent.” 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。

このようにして求められた候補範疇は、CKY 法を使用して求められたことから網羅的でかつ制約 1. を満たすことが保証され、また統語解析結果と文の型 S との単一化が成功した際の変数 x の値であることから制約 2. を満たすことが保証される。

以上のようにして最低限の制約を満たす統語範疇を考えると、解の候補は数多く存在する。しかし、単語の統語範疇として適切であると言語学者が判断するものは実際にはその中のごく一部であり、そこには他の制約が働いていると考えられる。

4. 制約を追加した解析

4.1 追加した制約

本稿では、解候補を絞り込む制約として以下を追加した場合の結果について考察する。

> 追加制約：
> 統語範疇 $T\backslash(T/NP)$、$T/(T\backslash NP)$（ただし T は変数）は、関数適用規則『$X/Y\ Y \Rightarrow X$』『$Y\ X\backslash Y \Rightarrow X$』の Y にはならない。

前節のパーザにおいて、関数適用規則『$X/Y\ \ Y\ \ \Rightarrow\ \ X$』『$Y\ X\backslash Y \Rightarrow X$』を適用する場合の条件として『$Y$ にあたる統語範疇が $T\backslash(T/NP)$、$T/(T\backslash NP)$ のいずれとも単一化不可能であること』を追加することでこの制約を取り入れている。

上記の追加制約は、該当するような振る舞いが実際の言語現象に見られないという言語学者のポリシーの 1 つを明示化したものである。この制約を満たさないものが言語学的に不適切であると主張するものでは必ずしもなく、別の文法を思い描けばそこには別のポリシーが存在するはずである。ここでは、[Steedman 96] による統語範疇を生成するような制約を設け、そこでのポリシーを明示化することを試みた。本来ならばもう一歩踏み込み、上記の制約を満たす統語範疇を考える文法がそうでない文法と比較して言語学的にどのような意味をもつのかということについても考察を与えるべきところであるが、別の機会に譲る。

4.2 制約を追加した解析結果

単語 X の統語範疇が未知である文 “John X Vincent.” について、追加制約を導入した条件で統語範疇の候補を出力し、追加前に出力された候補と比較した。解析時に用いた辞書は以下の通りである。

John: $T/(T\backslash NP)$, $T\backslash(T/NP)$
Vincent: $T/(T\backslash NP)$, $T\backslash(T/NP)$

制約追加前に生成された候補範疇数が 18 であったのに対し、制約を追加したことにより以下の 2 統語範疇に減少した。

> 未知語 X の統語範疇の候補：
> A. $(S\backslash NP)/NP$
> B. $(S/NP)\backslash NP$

このとき、制約追加により候補から除外された統語範疇は以下のようであった。

$(S/(T/(T\backslash NP)))\backslash(T\backslash(T/NP))$	$(S/(T/(T\backslash NP)))\backslash(T/(T\backslash NP))$
$(S/(T\backslash(T/NP)))\backslash(T\backslash(T/NP))$	$(S/(T\backslash(T/NP)))\backslash(T/(T\backslash NP))$
$(S\backslash(T\backslash(T/NP)))\backslash(T\backslash(T/NP))$	$(S\backslash(T\backslash(T/NP)))\backslash(T/(T\backslash NP))$
$(S\backslash(T/(T\backslash NP)))\backslash(T\backslash(T/NP))$	$(S\backslash(T/(T\backslash NP)))\backslash(T/(T\backslash NP))$
$(S/(T/(T\backslash NP)))\backslash NP$	$(S/(T\backslash(T/NP)))\backslash NP$
$(S\backslash(T\backslash(T/NP)))/NP$	$(S\backslash(T/(T\backslash NP)))/NP$
$(S\backslash NP)/(T\backslash(T/NP))$	$(S\backslash NP)/(T/(T\backslash NP))$
$(S/NP)\backslash(T\backslash(T/NP))$	$(S/NP)\backslash(T/(T\backslash NP))$

また、文 “Peter saw John with a telescope.” において with を未知語とし、以下の辞書項目を用意して解候補を求めたところ、制約追加前に生成されたの候補範疇数が 549 であったのに対し、制約追加後は 143 に減少した。

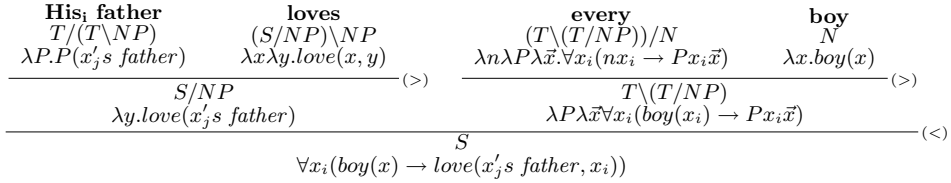

図 1: *His$_i$ father loves [every boy]$_i$. の導出

Peter:	$T/(T\backslash NP)$,	$T\backslash(T/NP)$
saw:	$(S\backslash NP)/NP$	
John:	$T/(T\backslash NP)$,	$T\backslash(T/NP)$
a:	$(T/(T\backslash NP))/N$,	$(T\backslash(T/NP))/N$
telescope:	N	

4.3 解析結果についての考察

文 “John X Vincent.” の解析で解候補として出力された 2 つの統語範疇のうち、言語学的に適切な統語範疇は A. であると考えられる。B. の統語範疇を許した場合、『*His$_i$ father loves [every boy]$_i$.』について図 1 のような導出が可能となるが、導出結果 S の意味表示より、$i = j$ だとすると本来なら許されないはずの参照が可能となってしまう。

このような不適切な予測が出ることから、上記の 2 つの統語範疇のうち B. を解候補から除外しなければならないが、どのような制約を用いることが可能であるのかは現時点では明らかではない。

5. 今後について

現時点における解析結果を分析する限り、望ましい統語範疇が解候補に含まれており、また候補から除外された統語範疇のうちにここで想定した言語学者のポリシーに沿ったものは含まれていなかった。これにより、今回加えた制約が我々の目指す CCG の統語範疇を生成するのに妥当であることが示唆されるが、品詞による例外の有無等を今後別の例文を使って検証していく必要がある。

また、より複雑で単語数の多い例文を扱う際には、ここで加えた制約だけでは解候補を絞り込みきれないことも明らかになった。このことは、他にも複数の制約が複雑に絡み合って作用している可能性を示唆している。今後、そのような複雑な例文について今回の制約を適用したときにどのような統語範疇が除外されきれずに残るかを分析し、その中から言語学者が思い描く統語範疇だけを切り出すためにどのような制約を与えることが可能かを検証していくことを考えている。

6. 本手法の言語学的意義

このようなアプローチをとることは、工学的な面のみならず、言語学的にも意義があることであると考えている。

CCG に関していえば、論理としての CCG の記述能力に従って生成されるもの全てが実際に自然言語の記述に使用されているわけではない。言語学者が無意識に対象としているのはその中のごく一部の領域のみであり、個々の言語学者はさらに各々のポリシーに基づいて領域の中の一部を用いて文法を構築していると考えられる。

これまで、語に与えた統語範疇の良し悪しは議論され得たが、その統語範疇に決定するに至ったポリシーについては明示化されていなかったため議論の対象になり得なかった。そのため、潜在的には数多く存在している候補の中で言語学者がある特定の統語範疇を選んだことの妥当性を裏付ける厳密な議論もされ得なかった。

言語学者のポリシーを明らかにしそれについて議論を可能にすることは、一歩踏み込んだ文法理論間の比較を可能にするだけでなく、自然言語が持っている共通の制約や性質、人間が言語の振る舞いから無意識に学習するものなどを解明することにつながると期待される。

7. 終わりに

本稿では、言語学者が語に統語範疇を与える時のポリシーを明示化し、制約として与えることで未知語の統語範疇を自動推定することを目標とし、その第一歩として「統語範疇 $T\backslash(T/NP)$、$T/(T\backslash NP)$（ただし T は変数）は、関数適用規則『$X/Y\ Y \Rightarrow X$』『$Y\ X\backslash Y \Rightarrow X$』の Y にはならない」という制約を用いて解の絞り込みを試みた。将来的には supertagging と同等以上の精度をもち、未知の統語範疇にも対応できる未知語解析器を実現したいと考えている。

参考文献

[Clark 07] Clark, S. and Curran, J. R.: Wide-coverage efficient statistical parsing with CCG and log-linear models, *Computational Linguistics*, Vol. 33, No. 4, pp. 493–552 (2007)

[Makino 97] Makino, T., Torisawa, K., and Tsujii, J.-I.: LiLFeS–practical programming language for typed feature structures, in *the Proceedings of Natural Language Pacific Rim Symposium (NLPRS)* (1997)

[Ninomiya 06] Ninomiya, T., Matsuzaki, T., Tsuruoka, Y., Miyao, Y., and Tsujii, J.: Extremely Lexicalized Models for Accurate and Fast HPSG Parsing, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 155–163, Sydney, Australia (2006), Association for Computational Linguistics

[Steedman 96] Steedman, M.: *Surface structure and interpretation*, Vol. 30, The MIT press, Cambridge, MA (1996)

[Yao 09] Yao, X., Ma, J., Duarte, S., and Çöltekin, Ç.: 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)

[尾崎 13] 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)