COLING2012 参加報告（その 4）
– ベストペーパーの紹介 –

原　一夫[†]

本稿では， Coling 2012 best paper （以下では (Nguyen et al. 2012) と表記する）の簡単な紹介を行う．

タイトル

Accurate Unbounded Dependency Recovery using Generalized Categorial Grammars

著者

Luan NGUYEN（ミネソタ大学)，Marten VAN SCHIJNDEL（オハイオ州立大学)，William SCHULER（オハイオ州立大学）

1　研究の背景（長距離依存関係の同定は難しい）

離れた単語間の依存関係（長距離依存関係）を正しく同定することは，情報抽出や質問応答など，文の意味理解を必要とするタスクを解決する上で重要である．しかし，Penn Treebank スタイル (Bies et al. 1995) のアノテーションは，長距離依存の学習に適するとは言えない．実際，Penn Treebank を学習用データとして利用する（句構造解析に基づく）多くの構文解析器は，長距離依存の解析を苦手とする．

長距離依存を含む文の例として “*the person who officials say stole millions*” の単語間依存関

[†]情報システム研究機構・国立遺伝学研究所, National Institute of Genetics

図 1　長距離依存を含む文を例示する．依存関係を有向枝で示す．太い枝は長距離依存に関連する枝である．

係を図 1 に示す[1]．挿入句（または，関係詞節を埋め込む節）*“officials say”* が存在するために，関係詞節の動詞 *“stole”* とその目的語 *“person”* は互いにやや離れた位置にある．

一方，この例文に対する Penn Treebank スタイルのアノテーション，すなわち句構造解析木を 図 2(a) に示す．ここで，句構造文法が定めるカテゴリ（S, NP, VP など）を用いて構文木を作ろうとすると，空要素（-NONE-）を導入する必要があることに注意する．

他方，組合せ範疇文法 (CCG) (Steedman 2000; Clark and Curran 2007) は，図 2(c) に見るように，空要素を導入することなく長距離依存を説明できる．しかし，カテゴリ（NP/NP, (S\NP)/NP, (NP\NP)/(S/NP) など）の数が非常に多くなるため，CCG パーザの学習には多数の訓練データを要すると考えられる．

2　提案手法

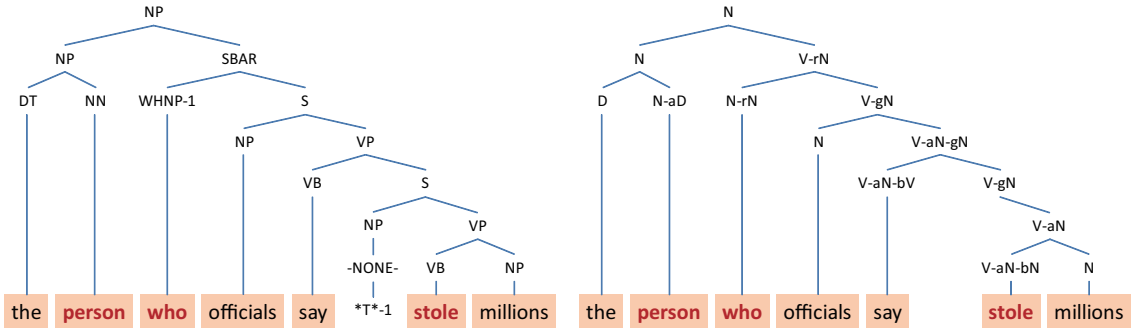
本稿が紹介する Coling 2012 best paper (Nguyen et al. 2012) は，CCG の長距離依存を解析する上での長所を保持しつつ，カテゴリ数を減らした文法を提案する．以下に提案文法の特徴を述べる．

まず，提案文法は，少数の基本カテゴリ（primitive category type）として，“D” を冠詞，“N” を名詞句，“V” を動詞句，などのように定義する．そして，基本カテゴリを type-constructing operator を用いて組み合わせ，新たなカテゴリを作る．例えば，主語の欠けた動詞句は，type-constructing operator “-a” を用いて “V” を “N” と結合したカテゴリ “V-aN” として表される．同様に，目的語の欠けた動詞句は “-b” を用いて “V” を “N” と結合したカテゴリ “V-bN” として表される．さらに，“-g” を用いて空要素を含む動詞句を表すカテゴリ “V-gN” が作成され，“-r” を用いて関係代名詞を表すカテゴリ “N-rN” が作成される．

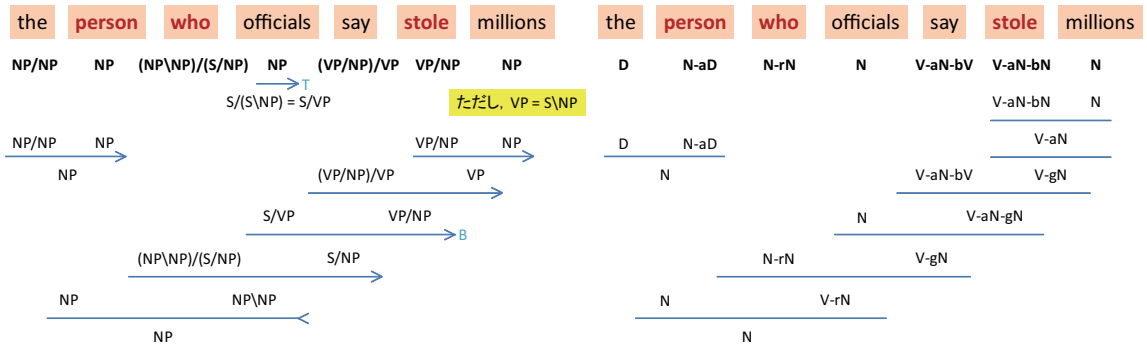
また，提案文法は語彙化文法である．すなわち，解析の際には，解析対象文に現れる全ての単語に対して上記のようにして作られたカテゴリが割り当てられる．そして，推論規則[2]を順次適用することにより，解析木が導出される．提案文法による解析例を図 2(d) に示す．

[1] オリジナル論文 (Nguyen et al. 2012, Figure 1) では，この文に対する predicate-argument dependency のグラフ表現が図示されている．

[2] 推論規則は各々のカテゴリに関連付けられる関数の合成方法を規定する．具体的な推論規則はオリジナル論文を参照されたい．



(a) Penn Treebank スタイルの構文木　　(b) 提案文法による構文木



(c) 組合せ範疇文法（CCG）による解析　　(d) 提案文法による解析

図 2　図 1 の例文に対して，Penn Treebank スタイルの構文木と提案文法による構文木，および，CCG による解析と提案文法による解析を並べて掲示する．

なお，Penn Treebank スタイルの構文木（図 2(a)）に対し，sed コマンドに似たスクリプトを適用することで，提案文法の構造木（図 2(b)）に容易に変換できることを，著者らは提案手法の利点として挙げている．

3　既存手法との比較

(Rimell et al. 2009) が作成した長距離依存同定（Unbounded Dependency Recovery）タスクのデータセットを用い，提案手法を既存の構文解析器と比較するために行われた実験の結果を表 1 に掲げる．ここで，長距離依存同定タスクは，以下の 7 つのサブタスクから成る．

Obj RC　関係詞節の目的語に関するもの．例：*the paper which I wrote*

Obj Red　関係代名詞が省略された関係詞節の目的語に関するもの．例：*the paper I wrote*

Sbj RC　関係詞節の主語に関するもの．例：*the instrument that measures depth*

表 1　長距離依存同定の実験の結果（精度）を示す．提案手法，Malt，MST 以外の精度は (Rimell et al. 2009) の再掲である．

比較手法	Obj RC	Obj Red	Sbj RC	Free	Obj Q	RNR	Sbj Embed	Total
Enju(Miyao and Tsujii 2005)	47.3	65.9	82.1	**76.2**	32.5	47.1	32.9	54.4
C&C(Clark and Curran 2007)	**59.3**	62.6	80.0	72.6	27.5	**49.4**	22.4	53.6
Malt(Nivre et al. 2006)	40.7	50.5	**84.2**	70.2	16.2	39.7	23.5	46.4
MST(McDonald 2006)	34.1	47.3	78.9	65.5	18.8	45.4	37.6	46.1
Stanford(Klein and Manning 2003)	22.0	1.1	74.7	64.3	41.2	45.4	10.6	38.1
DCU(Cahill et al. 2004)	23.1	41.8	56.8	46.4	27.5	40.8	5.9	35.7
Rasp(Briscoe et al. 2006)	16.5	1.1	53.7	17.9	27.5	34.5	15.3	25.3
提案手法 (Nguyen et al. 2012)	52.7	**69.2**	68.4	69.0	**57.5**	26.4	**38.8**	**54.6**

Free　先行詞のない関係詞節に関するもの．例：*I heard what she said*

Obj Q　疑問文の目的語に関するもの．例：*What did you eat?*

RNR　右方節点繰上げに関するもの．例：*Mary saw and Susan bought the book*

Sbj Embed　埋め込まれた関係詞節の主語に関するもの．例：*the responsibility which the government said lay with the voters*

表 1 によれば，提案手法は Enju（HPSG）および C&C（CCG）と同程度かやや優れた精度を達成しているように見える．しかし，著者らに問い合わせたところ，Enju，C&C の精度を測った Rimell らは再現の難しい手順で実験を行っており，しかも実験データを保存されておらず，よって，精度の違いの元となる原因の詳細な分析はできなかったとのことである．また，RNR に関して提案手法の精度が落ちているのは，訓練データを提案文法の構文木に変換するときに生じたエラーのためと考えられるとのことである．なお，CCG と比して提案文法がどの程度カテゴリ数を減らしたかについても，調べられなかったとのことである．

4　コントリビューション

以上をまとめると，(Nguyen et al. 2012) のコントリビューションは，次の通りである．

(1)　長距離依存を無理なく解析できる CCG の特長を保持しながら，カテゴリ数を減らした新しい文法を提案した．

(2)　Penn Treebank スタイルの構文木を提案文法の構文木に手軽に変換できるコンバータを作成した[3]．

(3)　(Rimell et al. 2009) の長距離依存同定タスクの実験を行い，提案文法を評価した．

[3]変換スクリプト等の所在は http://sourceforge.net/projects/modelblocks である．

謝辞

本稿を執筆する機会を与えていただきました言語処理学会に感謝いたします．また，本稿を執筆する上で，奈良先端科学技術大学院大学・自然言語処理学研究室の夏の集中勉強会（2010年度）で行われた CCG に関する議論は大いに役に立ちました．当時の勉強会参加者の皆さまに深く感謝いたします．

参考文献

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M. A., and Schasberger, B. (1995). "Bracketing Guidelines for Treebank II Style Penn Treebank Project." Tech. rep., University of Pennsylvania.

Briscoe, T., Carroll, J., and Watson, R. (2006). "The Second Release of the RASP System." In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 77–80 Sydney, Australia. Association for Computational Linguistics.

Cahill, A., Burke, M., O'Donovan, R., Van Genabith, J., and Way, A. (2004). "Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations." In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 319–326 Barcelona, Spain.

Clark, S. and Curran, J. R. (2007). "Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models.." *Computational Linguistics*, **33** (4), pp. 493–552.

Klein, D. and Manning, C. D. (2003). "Accurate Unlexicalized Parsing." In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430 Sapporo, Japan. Association for Computational Linguistics.

McDonald, R. (2006). *Discriminative Learning and Spanning Tree Algorithms for Dependency Parsing*. Ph.D. thesis, University of Pennsylvania.

Miyao, Y. and Tsujii, J. (2005). "Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 83–90 Ann Arbor, Michigan. Association for Computational Linguistics.

Nguyen, L., Van Schijndel, M., and Schuler, W. (2012). "Accurate Unbounded Dependency Recovery using Generalized Categorial Grammars." In *Proceedings of COLING 2012*, pp. 2125–2140 Mumbai, India. The COLING 2012 Organizing Committee.

Nivre, J., Hall, J., and Nilsson, J. (2006). "MaltParser: a data-driven parser-generator for

dependency parsing." In *Proceedings of LREC-2006*.

Rimell, L., Clark, S., and Steedman, M. (2009). "Unbounded dependency recovery for parser evaluation." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pp. 813–821 Stroudsburg, PA, USA. Association for Computational Linguistics.

Steedman, M. (2000). *The syntactic process*. MIT Press, Cambridge, MA, USA.

略歴

原 一夫：1994 年東京大学工学部計数工学科卒業．1996 年同大学院工学系研究科修士課程計数工学専攻修了．（株）ドラゴンジェノミクス（現・タカラバイオ），三共株式会社（現・第一三共）勤務を経て，2008 年奈良先端科学技術大学院大学情報科学研究科（自然言語処理学）博士後期課程修了．現在国立遺伝学研究所研究員．博士 (工学)．

（2012 年 11 月 30 日依頼）
（2013 年 1 月 21 日受付）