Acknowledgment

This work was supported by a Grant-in-Aid for Research on Applied Use of Statistics and Information, Health and Labour Sciences Research and Clinical Research for Development of Preventive Medicine and New Therapeutics from the Ministry of Health, Labour and Welfare.

References

[1] WHO Medicines, "Traditional and Complementary Medicine," http://www.who.int/medicines/areas/traditional/en/.

[2] Declaration of Alma-Ata International Conference on Primary Health Care, Alma-Ata, USSR, September 1978, http://www.who.int/publications/almaata_declaration_en.pdf.

[3] P.-F. Gao and K. Watanabe, "Introduction of the World Health Organization project of the International Classification of Traditional Medicine," *Journal of Chinese Integrative Medicine*, vol. 9, no. 11, pp. 1161–1164, 2011.

[4] K. Watanabe, X. Zhang, and S.-H. Choi, "Asian medicine: a way to compare data," *Nature*, vol. 482, no. 7384, p. 162, 2012.

[5] "ICD11 beta," http://apps.who.int/classifications/icd11/browse/f/en.

[6] G. S. de Morant, *Chinese Acupuncture*, Paradigm Publication, Tokyo, Japan, 1994.

[7] G. A. Plotnikoff, K. Watanabe, and F. Yashiro, "Kampo-from old wisdom comes new knowledge," *Herbal Gram*, vol. 78, pp. 46–57, 2008.

[8] K. Terasawa, "Evidence-based reconstruction of Kampo medicine: part I—is Kampo CAM?" *Evidence-Based Complementary and Alternative Medicine*, vol. 1, no. 1, pp. 11–16, 2004.

[9] K. Watanabe, K. Matsuura, P. Gao et al., "Traditional Japanese Kampo medicine: clinical research between modernity and traditional medicine—the state of research and methodological suggestions for the future," *Evidence-Based Complementary and Alternative Medicine*, vol. 2011, Article ID 513842, 19 pages, 2011.

[10] E. C. Moschik, C. Mercado, T. Yoshino, K. Matsuura, and K. Watanabe, "Usage and attitudes of physicians in Japan concerning traditional Japanese medicine (Kampo medicine): a descriptive evaluation of a representative questionnaire-based survey," *Evidence-based Complementary and Alternative Medicine*, vol. 2012, Article ID 139818, 13 pages, 2012.

[11] A. Ito, K. Munakata, Y. Imazu, and K. Watanabe, "First nationwide attitude survey of Japanese physicians on the use of traditional Japanese medicine (Kampo) in cancer treatment," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 957082, 8 pages, 2012.

[12] G. A. Plotnikoff and K. Watanabe, "New insights on women's health from Japan," *Minnesota Physician*, vol. 12, pp. 32–33, 2004.

[13] V. Scheid, T. Ward, W.-S. Cha, K. Watanabe, and X. Liao, "The treatment of menopausal symptoms by traditional East Asian medicines: review and perspectives," *Maturitas*, vol. 66, no. 2, pp. 111–130, 2010.

[14] Y. Gepshtein, G. A. Plotnikoff, and K. Watanabe, "Kampo in women's health: Japan's traditional approach to premenstrual symptoms," *Journal of Alternative and Complementary Medicine*, vol. 14, no. 4, pp. 427–435, 2008.

[15] Y. Sahashi, "Herbs covered by health insurance in Japan," *The Journal of Kampo, Acupuncuture and Integrative Medicine*, vol. 1, pp. 70–84, 2005.

[16] F. Yu, T. Takahashi, J. Moriya et al., "Traditional Chinese medicine and kampo: a review from the distant past for the future," *Journal of International Medical Research*, vol. 34, no. 3, pp. 231–239, 2006.

[17] S. Cameron, H. Reissenweber, and K. Watanabe, "Asian medicine: Japan's paradigm," *Nature*, vol. 482, no. 7383, p. 35, 2012.

[18] K. Terasawa, "Evidence-based reconstruction of Kampo medicine: part II— the concept of Sho," *Evidence-Based Complementary and Alternative Medicine*, vol. 1, no. 2, pp. 119–123, 2004.

[19] T. Wada, *Dosui Sagen*, K. Hayashi, Tokyo, Japan, 1805, (Japanese).

[20] K. Tatsuno, "Kyo-jitsu-ron (1)," *Journl of Kampo Medicine*, vol. 1, pp. 383–392, 1954 (Japanese).

[21] K. Tatsuno, "Kyo-jitsu-ron (2)," *Journl of Kampo Medicine*, vol. 1, pp. 445–457, 1954 (Japanese).

[22] E. Kaibara, *Yojokun*, Kodansha, Tokyo, Japan, 1982, Translated to modern Japanese by T. Ito.

[23] Japan Kampo Medicines Manufactures Association, http://www.nikkankyo.org/.

テキスト情報の解析と可視化に向けた高被覆日本語CCG構文解析の実現

植松 すみれ (東京大学)○，美馬 秀樹 (東京大学)

Inducing Japanese CCG Resources for Analysis of Information Encoded in Texts

Sumire UEMATSU and Hideki MIMA

ABSTRACT

This paper reports development of a Japanese CCG parser, which can be used as a building block for extracting information encoded in texts. Our method enables applying corpus-oriented grammar development to Japanese by integrating multiple dependency-based corpora. Quantitative evaluation of the method is presented in terms of the coverage of the obtained grammar and the accuracy of the parser. Necessary steps to use the parser for information extraction is also discussed in this paper.

Keywords: Natural Language Processing, CCG, Grammar development, Parsing

1 はじめに

本稿ではテキスト内の情報を解析する基礎技術としての日本語のCCG文法に基づく解析とその実現手法を紹介する。現在ビッグデータ解析の対象はセンサーデータ等が中心だが書物やWWW等のテキストには自然言語でかかれた膨大な知識が存在するはずである。そのような知識の自動抽出がテキスト解析技術によって実現すれば，より広範囲かつ抽象的な情報に対しても分析が可能になると考えられる。CCG構文解析は「深い解析」の一種で，現在日本語の解析として広く行われている係り受け解析 [7] と比べて述語項構造を同時に計算できる利点がある。

我々は複数の係り受けコーパスを組み合わせてCCG解析の擬似正解データをつくる手法を提案し，英語等で深い解析器構築に実績のあるコーパス指向文法開発手法と合わせて日本語CCG構文解析器を構築した [8]. この手法を紹介し実際にコーパスに適用した定量的評価の最新結果とともに，現状で抽出可能な情報について述べ情報抽出などの応用に向けた課題を示す。

2 情報抽出と構文解析

文の構造と意味を解析する構文解析は，自然言語処理の分野では従来から研究されてきたが，情報抽出などへの応用を考えた場合，言語の様々なレベルの言い換えを処理する機能が重要となる。言語の言い換えには，「ゆれ」「揺れ」等の表記ゆれから，受け身の格交替などの統語的な変化，文単位の言い換え等より大規模なものが存在するがそれらを吸収して意味を表示することが重要である。このうち統語的な言い換えを構文解析では処理することとなる。

語彙化文法を実世界のテキストに適用し処理に用いる場合，解析の精度に加えて，辞書の被覆率，つまり現実に使われている文に対してどの程度解析が可能か，という点が重要となる。コーパス指向文法開発手法は，木構造コーパスを変換し語彙化文法での構文木コーパスとすることで，十分な精度と辞書の被覆率を備えた解析器の実現に英語等において成功している [4]。

3 背景

3.1 CCGによる日本語解析

CCGは語彙化文法理論の一種であり，各単語に対する詳細な辞書情報と少数の組み合わせ規則を併せて言語のふるまいを説明する理論である。CCGの場合，統語的ふるまいはカテゴリと呼ばれる記法 (図1の $S\backslash NP$ など) で指定され，意味論上の役目はラムダ式を用いて表される (図1では省略)．文の解析では単語にカテゴリを割り当て，組み合わせ規則に従って句を

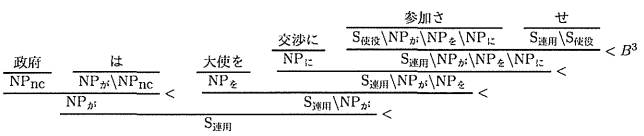

Fig.1 A simplified CCG analysis of the sentence 「政府は大使を交渉に参加させ (た)」.

構成しながらカテゴリの合成・ラムダ式操作を行なう事で，文の統語構造と意味表現が得られる．

今回は CCG 理論に基づいた日本語文法論 [9](以下戸次文法と呼ぶ) を目標とすべき文法の基礎とした。ただし，今回は戸次文法の説明する全ての現象を扱うわけではなく，用言のかき混ぜ，受け身，使役など重要かつ頻出する現象を正確に扱い，他の部分については簡略化した文法を目標として設定した．意味表現についてもラムダ式ではなく，用言について簡単な述語項構造を定義し深層格[*1]を用いている。

3.2 コーパス指向文法開発の日本語適用

コーパス指向文法開発と呼ばれるアプローチを適用して日本語解析器を構築した．コーパス指向文法開発は図 2 に示したように，構文情報がついたコーパスから目標文法での正しい導出木コーパスを作ることで，辞書と曖昧性解消モデルの学習データを得るアプローチである．元のコーパス内の情報から正しい導出木を構成できれば，その葉ノードの ⟨ 単語, カテゴリ ⟩ の組から辞書をつくる事ができる．また曖昧性解消モデルは，導出木コーパスでの導出木を正例，同じ文に対して獲得した辞書 (と組み合わせ規則) から導出可能な導出木を負例として学習することできる．

語彙化文法を獲得する元となるデータとして，同一文に対する統語構造と意味構造のアノテーションが必要となる．本研究では，日本語に対する大規模な統語構造つきコーパスとして現在も活発な付加アノテーションが行われている京大コーパス [5] を用いた。京大コーパスは統語情報として形態素解析，文節境界および文節間の係り受け関係のアノテーションを含み，同一テキストに対するアノテーションとして，NAIST テキストコーパス [10] による述語項構造および照応関係の情報，また，「と」コーパス [3] による助詞「と」でマークされた項と述語の関係の情報がある．本研究では上記

[*1] 動詞を基本形に直した際のガヲニト格

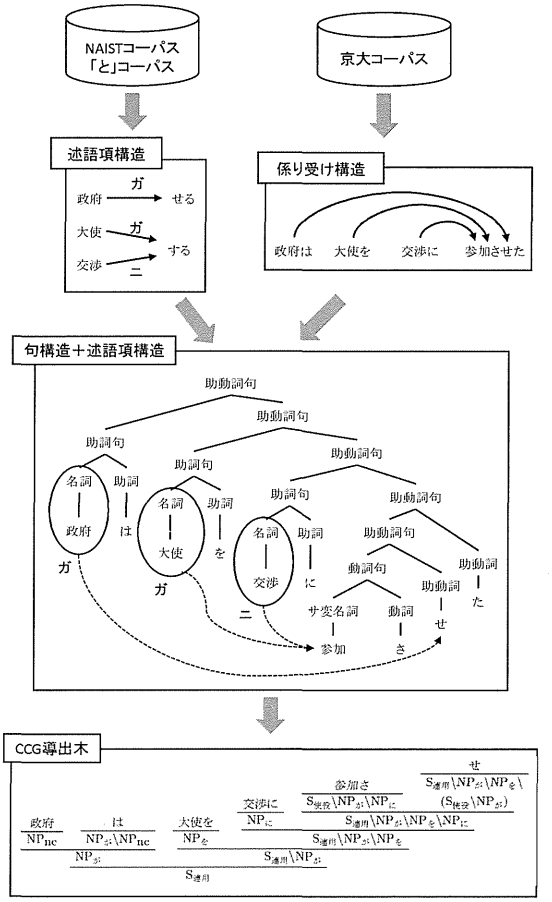

Fig.2 Conversion from dependency relations to a CCG derivation.

3 コーパスから得られる情報を統合的に利用する．ただし目標となる CCG 導出木を得るに当たって，上記 3 コーパスからは読み取れない統語・意味構造は，発見的ルールによる情報付加，あるいは人手による付加的アノテーションなど，何らかの方法で補う必要がある．

4 コーパス変換

3.2 節で述べたように，我々は京大コーパスの係り受け構造，NAIST コーパスの述語項構造などを統合して CCG 導出木へ変換し，日本語 CCG 文法の辞書を獲得した。コーパス変換処理では，まず目標とする CCG 文法に依存しない中間表現として句構造データを想定し，まず京大コーパス並びに関連コーパスから，それらのアノテーションを統合した句構造データへの変換を行い，その後句構造データから目標 CCG 文法の導出木への変換，の 2 段階で変換を行うこととした (図 2).

	総文数	変換文数	変換成功率 (%)
第 1 段階目	24,283	24,116	99.3
第 2 段階目	24,116	22,852	94.8

Table1 Statistics of corpus conversion.

	LP	LR	LF	UP	UR	UF	Lex
開発	85.1	84.7	84.9	90.9	90.5	90.7	92.3
テスト	85.2	84.6	84.9	91.0	90.4	90.7	92.3

Table2 LP, LR, and LF refer to labeled precision, recall, and F-score, respectively. UP, UR, and UF are for unlabeled.

■係り受け構造から句構造へ 大きく分けて (1) 文節内の構造の認定 (2) 文節間の係り受け関係から句間の関係への変換 の 2 段階からなる．(1) では文節内の形態素列を 2 分木としてまとめ上げるため，原則 (例えば用言を中心とした文節の場合は，全体が右上がりの木構造) と併せて例外的な構造をとるケースを CFG ルールとして表したうえで，各文節について決定的にルール適用を行った．(2) の句構造への変換処理は，(1) で決定した各文節に対する部分木を，係り受け構造の通りに組み合わせることで行う．この際，係り側の句の部分木を，受け側の部分木のどの位置に接合するかを決定する必要があるが，これも係り側の句の種類，最右の形態素の活用形，京都コーパスの係り受け関係ラベル (並列句か否か) などを基にしたルールに従って行った．

■句構造から導出木へ 原則として句構造と木の形が同じ導出木を想定し，述語項構造アノテーション等から導出木に変更を加えていく方法をとった．具体的には (1) 導出木ノードに局所制約をかける (例えば文法上，カテゴリの形が定まっている助動詞のノードには $S\backslash S$ のカテゴリを与える等)，(2) 導出木の各枝分かれについて，親子ノードのラベル，カテゴリなどから組み合わせ規則を決定する，(3) 根ノードに S(文を表すカテゴリ) を割り当てた上で (2) で指定された組み合わせ規則を適用して局所制約を統合し導出木を獲得する，の 3 段階で変換を行った．

■導出木からの辞書獲得 最後に葉ノードの形態素とカテゴリの組を辞書に登録した。特に動詞の葉ノードの場合，活用形，態などによりカテゴリに受け身接続形か否かなどの素性を与え，活用形ごとに辞書エントリを構成した。またかき混ぜを含んだり，ガ格の省略がある動詞の場合，標準形カテゴリ (ガ格があり，かつヲ，ニ，ト格の順に項がならぶ) を想定，得られたカテゴリは標準型からの派生カテゴリとして辞書に登録した．

5 結果

京大コーパス Version 4.0，NAIST コーパス Version 1.5，「と」コーパス Version 1.0 を用いて文法を抽出，獲得文法による解析実験を行った．京大コーパスを用いた係り受け解析の実験 [7] と同様にデータを分割し，訓練セット，開発セット，テストセットを用意した．

5.1 コーパス変換と辞書抽出

文法開発でのコーパス変換については，京大コーパスおよび関連コーパスから句構造への変換，句構造から CCG 導出木への変換を行った結果，訓練セット (全 24,283 文) から 22,852 の導出木が得られ，94.1% の文を導出木へ変換することができた．表 1 は変換の各段階における成功率であり，第 2 段階の句構造から CCG 導出木への変換において，制約の相互矛盾によって文法を満たさない導出木が検出されていることがわかる．

次に導出木の葉ノードから ⟨ 形態素基本形, 品詞, 活用形, カテゴリ ⟩ の組を抽出し，辞書を構成した．導出木コーパス (総語数は 616,305 語，カテゴリ種総数 699) から得られた辞書エントリ，つまり上記の 4 つ組の種類数は 84,620 個となった。

5.2 辞書の被覆率とパーザの解析精度

辞書被覆率とは，未知文中の形態素に対して正しいカテゴリが辞書に登録されている割合を示すものである．今回は訓練セットから抽出した辞書の被覆率を開発セット，テストセットに対して測定した．開発セットの 100 語以下の文に対して被覆率を測った結果，この文法の辞書による被覆率は 99.4% であり高被覆な辞書が得られたといえる．

訓練セットから得られた文法を使ってテストセットの文を解析した結果の精度を表 2 に示す．解析には [6] で用いられた語彙化文法用パーザと曖昧性解消モデルを用い，入力は正解形態素付きの文とした．またテスト文に対する正解としては，テストセットに 4 節の変換を行って得られた導出木を正解の導出木として扱った．

今回の解析精度と英語における CCG パーザの精度 [2] とは，言語や実装の違いもあり直接比べられるものではない．ただし我々の日本語文法の被覆率は英語 CCG 文法の被覆率 (99.63%) とほぼ同じレベルであるのに対し，単語へのカテゴリ割り当て正解率と解析精度

は英語の場合と比べて 2〜3 ポイント低くなっている単語単位，構造単位で曖昧性解消での課題が示唆された．

6 応用に向けた課題

2 節で述べた情報抽出への応用から見た評価を述べる。獲得した CCG 辞書は形態素単位の被覆率が 99.4% と実世界テキストの解析に十分な被覆率であると考えられる。ただし曖昧性解消も含めた解析精度は 5 節で述べたように英語パーザと比べて精度の落ち込みが見られ，日本語解析で一般的な文節係り受けとの精度比較[*2]も，[7] によるとラベルなし単語間係り受け精度が 95.1% となっており，これと比べて表 2 の UP とは差がある。解析器のエラーは後処理にも伝搬するため，情報抽出等後処理に利用することを考えると精度の向上が望ましい。曖昧性解消モデルに係り受け解析で用いられる素性を採用するなど日本語への適応を行うことで向上を図ることが考えられる。

また今回獲得した文法は受け身，使役などを扱えるため，情報抽出を行う上で重要になる用言の深層格を認識，出力することができる。ただしこの文法は第 1 段階として実装したものであり，さらに並列構造など他の重要な現象の扱いを改善する必要がある。また完全な意味表現を出力するためには各単語に対してラムダ式を割り当てる辞書が必要であり，Bos[1] のような辞書獲得の手法の開発が必要である。

また本手法の課題として変換して得られた導出木の正確性の評価と向上が挙げられる。コーパス変換はパターンルールによる自動的変換であるため，例外的な構造をとる文に対して不適切なルール適用が行われる可能性がある。変換の各段階でルールを詳細化することで，より精緻で正確な分析が可能になるが，複合名詞の内部構造の決定等は本質的に人手によるアノテーションを必要とするため，木構造に対するアノテーションの追加，修正を行うことも検討中である。

7 まとめ

知識の自動抽出を目指したテキスト内の情報を解析する基礎技術としての日本語の CCG 文法に基づく解析とその実現手法を紹介した。複数の係り受けコーパスを組み合わせることで CCG 解析の擬似正解データをつくる手法と，コーパス指向文法開発手法の日本語への適用を可能にし，日本語 CCG 構文解析器を構築結果を紹介した。日本語構文解析研究において広く用いられる京都大学テキストコーパスとその関連コーパスに対して，提案手法を適用した結果は，94.1% の文の変換に成功し獲得辞書の被覆率は 99.4% であった。獲得文法を用いた解析の精度は英語 CCG パーザや日本語係り受けの精度と比較してやや低く曖昧性解消モデルの改善に必要性が示唆される。

現在用言に関する深層格情報の出力が可能であるが，情報抽出など応用を考えた場合，ラムダ式による意味表現辞書の開発，変換後導出木の正確性の評価とその向上のための追加アノテーションが今後の課題となる。

[*2] 依存関係の単位の違い，文節係り受けと CCG の取り扱う依存関係の範囲の違いから単純に比較できるものではない

参考文献

[1] Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. Wide-coverage semantic representations from a CCG parser. In *Procc of COLING*, pp. 1240–1246, 2004.

[2] Stephen Clark and James R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Comp. Ling.*, Vol. 33, No. 4, 2007.

[3] Hiroki Hanaoka, Hideki Mima, and Jun'ichi Tsujii. A Japanese particle corpus built by example-based annotation. In *Proc. of LREC*, 2010.

[4] Julia Hockenmaier and Mark Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Comp. Ling.*, Vol. 33, No. 3, pp. 355–396, 2007.

[5] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proc. of LREC*, 1998.

[6] Yusuke Miyao and Jun'ichi Tsujii. Feature forest models for probabilistic HPSG parsing. *Comp. Ling.*, Vol. 34, No. 1, pp. 35–80, 2008.

[7] Manabu Sassano and Sadao Kurohashi. A unified single scan algorithm for Japanese base phrase chunking and dependency parsing. In *Proc. of ACL-IJCNLP*, 2009.

[8] 植松すみれ, 松崎拓也, 花岡洋輝, 宮尾祐介, 美馬秀樹. 統語・意味コーパスの統合と再解釈による大規模な日本語 CCG 文法の開発. 人工知能学会全国大会, 2013.

[9] 戸次大介. 日本語文法の形式理論—活用体系・統語構造・意味合成—. 日本語研究叢書. くろしお出版, 2010.

[10] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション:NAIST テキストコーパス構築の経験から. 自然言語処理, Vol. 17, No. 2, pp. 25–50, 2010.