2D5-OS-03b-4

意味表現を所与とする自然言語表現の生成モデル

Generative Model of Natural Language Expressions for given Semantic Representations

麻生 英樹[*1]
Hideki ASOH

[*1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology, AIST

Modeling generative process of natural language is an important problem of natural language processing. Although many probabilistic models have been proposed so far, most of them are in surface structure level such as n-gram model, or in surface structure and syntax structure levels such as probabilistic CFG. Models include semantic structure level have not been investigated extensively. In this presentation, first an overview of two existing generative models of natural language expression from a given semantic representation is introduced. After that, such models and our method of natural language generation from semantic representation are compared. The way to coordinate them is also investigated.

1. はじめに

自然言語を用いて人間と円滑にコミュニケートできる知的システムを実現するためには，多様な自然言語の表層表現と，推論や問題解決を可能にする意味表現とを相互に効率良く変換する処理が必要になる．これまでに，変形生成文法を始めとして，数多くの自然言語文の生成モデルが提案されてきているが，それら多くは統語論のレベルで閉じており，意味表現をも含むものはそれほど多くはない．

主辞駆動句構造文法（Head-driven Phrase Structure Grammar: HPSG) [Pollard 94] や組合せ範疇文法（Combinatory Categorial Grammar: CCG）[Steedman 96] では，フォーマルな意味表現の構造と統語的構造とが並行的に扱われているが，いずれも主として構文解析に適用されてきており，言語生成を扱う研究はあまり多くはない．

自然言語処理の分野では，近年，semantic parsing という名称で，自然言語の表層文を論理式などの意味表現に変換する問題が盛んに研究されているが [Ge 05, Wong 07]，その逆の問題，すなわち，意味表現から自然言語文を生成する方向の研究もそれほど行われていないように思われる．

我々は，自然言語の表層表現と意味表現を相互に円滑に変換するための意味表現について検討し，[麻生 10, 野口 10a, 10b] では，概念間の依存関係をベースとした意味表現 [高木 87] を介して，視覚情報から多様な言語表現を生成することができることを示した．また，[麻生 11] では，意味表現に基づいて，言語の生成性を，意味表現の生成性と，単語列の生成性に分けて議論するための枠組みを提案し，[麻生 12] では，自然言語文の確率的な生成モデルを，意味構造 L の生成モデル $P(L)$ と与えられた意味表現から表層文 S を生成する生成モデル $P(S|L)$ に分けて構築する Bayesian linguistics を提唱した．

本発表では，型付λ式やそれと同等程度の表現力を持つ形式の意味表現から自然言語文を生成する具体的な手法についてさらに検討した結果について報告する．最初に，既存の研究の中から White らの CCG を用いた言語生成および Lu らの hybrid tree を用いた言語生成について概説し，その後で，我々が提案してきている意味表現形式との関係や適用可能性について検討する．

連絡先：麻生英樹，産業技術総合研究所知能システム研究部門，asoh@aist.go.jp

2. CCG に基づく表層文生成

組合せ範疇文法（Combinatory Categorial Grammar: CCG）は，古典的な範疇文法の拡張として，Ades と Steedman によって提案された [戸次 10]．CCG は語彙化文法の一種であり，音韻表示，統語範疇，意味表示の組合せから成る語彙の集合と，それらを組み合わせて文を生成するための少数の組合せ規則から成る．CCG の特徴の一つは，統語範疇の型に基づく統語的な語の組合せと，意味表示の組合せとを並行的に扱うことができる点である．

従来，CCG の意味表示には，型付λ式が使われることが多かったが，近年，Hybrid Logic という様相論理の拡張に基づく意味表現，Hybrid Dependency Logic Semantics (HDLS) も用いられ始めている [Baldridge 02]．White らは，CCG に基づき，HDLS で表される意味表現から表層文を生成する手法を提案して実装した [White 03, 07]．そこでは，HLDS をフラット化した意味表現を入力とし，CCG の文法を用いて Chart 法で自然言語を生成している．さらに，POS-タグなども複合した n-gram によって生成結果をランキングするなどの統計的機械翻訳の手法を導入して，Pen Tree Bank から派生したコーパスである CCGBank 上でのカバレッジを上げることも行われている．この言語生成システムは CCG に対する構文解析器，文章生成器，コーパスなどのリソースを集めた Open CCG にも含められている [OpenCCG]．White らはコーパスに基づいて人手で作成した文法を用いているが，[Kwiatkowski 10] では，Higher Order Unification を用いて，表層表現と意味表現のペアの集合から確率的 CCG の文法を獲得する方法も提案されている．

3. Hybrid Tree に基づく表層文生成

Lu らは hybrid tree という意味表現と表層表現を混合した木構造の内部表現形式を用いて，Semantic Parsing およびその逆の意味表現からの自然言語生成の両方を同じ枠組みで扱えることを示した [Lu 08, 09]．さらに，[Lu 11] では，変数を持たない意味表現に対する hybrid tree を，より一般的な型付λ式に拡張した hybrid λ tree を用いて，λ表現から自然言語表現を生成する確率的なモデルを提案している．

Hybrid tree とは，木構造の意味表現に表層表現を混合させた木構造で，内部ノードは意味表現の要素，木の葉は表層の句，リーフを左 depth-first でつなぐと自然言語文になっている．

Lu らは, Markov 性を仮定した hybrid tree の確率的な生成モデルを提案した. 生成モデルは, 以下のような確率パラメータを持つ:

- MR parameter: $\rho(m_j|m_i, k)$
  意味表現要素 m_j を m_i の k 番目の子として生成する確率
- Emission parameter: $\theta(t\,|m_i, \Lambda)$
  表層の句／意味表現要素 t を, 文脈 Λ において, m_i から生成する確率
- Pattern parameter: $\phi(r\,|m_i)$
  hybrid pattern r を m_i の子の生成に使う確率

ここで, hybrid pattern とは, 表層記号と意味表現の組合せパターンであり, 以下の 4 種類が考慮されている:

m → w
m → [w]Y[w]
m → [w]Y[w]Z[w]
m → [w]Z[w]Y[w]

w は自然言語の単語列で, [ ] は optional であることを示す. Y, Z は意味表現要素である. ある hybrid tree の生成確率は, 上記の確率パラメータの積で与えられる. また, これらのパラメータは EM アルゴリズムの一種を用いて表層表現と意味表現のペアの集合であるコーパスから推定することができる.

hybrid tree の生成モデルを用いれば, 木構造の意味表現が与えられた条件の下で, 様々な表層表現の条件付き確率を求める表層文生成の問題と, 逆に, 表層表現が与えられた条件の下で, 意味表現を求める semantic parsing の問題の両者を, 事後確率計算の問題として自然に解くことができる. しかし, 文生成に関しては強すぎる独立性の仮定のために, 自然な表層表現の生成が困難であることがわかった. そこで, Lu らは, hybrid tree の生成確率モデル $P(S, L)$ と, tree-based CRF に基づく識別確率モデル $P(S \mid L)$ とを組み合わせて, より自然な表層表現を得ることを提案している [Lu 09]. また, 変数を含まない意味表現を前提とした手法を, 型つきλ式による意味表現にも拡張し, より複雑な意味表現から自然言語表現を生成する方法も提案している [Lu 11].

4. 考察とまとめ

我々は, 同義表現の言い換えや視覚情報と言語情報の相互変換に適した意味表現として, 高木らが提唱した概念依存関係に基づく意味表現 [高木 87] に着目し, その意味表現上の同義変形と意味表現の単語分割によって表層文を生成する方法を提案してきた. その手法を用いて, 対話システムや視覚情報を言語化するシステムも構築している [麻生 10, 野口 10a, 10b].

そこで用いている意味表現は木構造で表現されているため, たとえば, Lu らの hybrid tree の考え方を適用することも可能であると考えられる. しかし, その場合, 以下のような点について検討する必要があると思われる:

- 同じ意味表現から複数の言い回しの表層文(同義文)が生成されることをどのように説明できるか？
- 意味表現中のどの要素を主節として, どの要素に言及するかをどのようにして制御できるか？

本発表においては, 上記の点に関して, White や Lu らの言語生成手法がどのように対応できるのかを検討するとともに, 我々の意味表現を用いた表層文生成手法に, hybrid tree のような確率モデル的な考え方をどのように融合できるかについても議論したい.

参考文献

[麻生 10] 麻生英樹, 野口靖浩, 高木朗, 小林一郎, 近藤真, 三宅芳雄, 岩橋直人, 伊東幸宏: 視覚情報から多様な言語表現を生成するための意味表現形式, 第 24 回人工知能学会全国大会, 2G1-OS3-7, 2010.

[麻生 11] 麻生英樹: 自然言語の生成性について, 第 25 回人工知能学会全国大会, 3H2-OS3-10, 2011.

[麻生 12] 麻生英樹: 確率モデルからの記号の創発－Bayesian Linguistics に向けて－, 人工知能学会誌, Vol.27, No. 6, 546-554, 2012.

[Baldridge 02] Baldridge, J. and Kruijff, G.-J. M.: Coupling CCG and hybrid logic dependency semantics, Proc. ACL-02, 319-326, 2002.

[戸次 10] 戸次大介: 日本語文法の形式理論, くろしお出版, 2010.

[Ge 05] Ge, R. and Mooney, R. J.: A statistical semantic parser that integrates syntax and semantics, Proc. CNLL-05, 9-16, 2005.

[Kwiatkowski 10] Kwiatkowski, T., Zettlemoyer, L., Goldwater, S. and Steedman, M.: Inducing probabilistic CCG grammars from logical form with higher-order unification, Proc. EMNLP-10, 2010

[Lu 08] Lu, W., H. T. Ng, W. S. Lee, and L. Zettlemoyer: A generative model for parsing natural language to meaning representations, Proc. EMNLP-08, 782-791, 2008.

[Lu 09] Lu, W., H. T. Ng, and W. S. Lee: Natural language generation with tree conditional random fields, Proc. EMNLP-09, 400-409, 2009.

[Lu 11] Lu, W. and Ng, H. T.: A probabilistic forest-to-string model for language generation from typed lambda calculus expressions, Proc. EMNLP-11, 1611-1622, 2011.

[野口 10a] 野口靖浩, 麻生英樹, 高木朗, 小林一郎, 三宅芳雄, 岩橋直人, 伊東 幸宏: 視覚情報から多様な言語表現を生成するための意味表現形式, 第 24 回人工知能学会全国大会 2G1-OS3-8, 2010.

[野口 10b] 野口靖浩, 麻生英樹, 高木朗, 小林一郎, 近藤真, 三宅芳雄, 岩橋直人, 伊東幸宏: 視覚情報から言語表現を生成するシステムの試作, 第 58 回言語・音声理解と対話処理研究会(SIG-SLUD)技術報告, pp.43-48, 2010.

[OpenCCG] OpenCCG: http://openccg.sourceforge.net

[Pollard 94] Pollard, C. and Sag, I. A.: Head-Driven Phrase Structure Grammar, University Chicago Press, 1994.

[Steedman 96] Steedman, M.: Surface Structure and Interpretation, MIT Press, 1996.

[高木 87] 高木朗, 伊東幸宏: 自然言語の処理, 丸善, 1987.

[White 03] White, M. and Baldridge, J.: Adapting chart realization to CCG, Proc. EWNLG-03, 2003.

[White 07] White, M., Rajkumar, R., and Martin, S.: Towards broad coverage surface realization with CCG, Proc. UCNLG+MT-07, 2007.

[Wong 07] Wong, Y. W. and Mooney, R. J.: Learning synchronous grammars for semantic parsing with lambda calculus, Proc. ACL-07, 960-967, 2007.