工学ジャーナル　第 11 巻
2012 年 3 月

主辞駆動句構造文法のための同期文法の実現に向けて

Towards Synchronous Head-driven Phrase Structure Grammars

二宮 崇*

Takashi NINOMIYA*

Abstract: This study proposes a model of synchronous Head-driven Phrase Structure Grammar for machine translation. Statistical machine translation currently receives much attention in the field of machine translation because of ability to learn its model from parallel corpora automatically. However, statistical machine translation suffers from difficulty in translating between language pairs which are structurally different, due to lack of grammatical hypothesis or models. Synchronous grammars provide grammatical relations of two languages, which enable machine translation to incorporate grammatical structures into the translation model. We developed an experimental synchronous HPSG grammar, and its performance was evaluated on English-Japanese machine translation as preliminary experiments.

Key words: Head-driven Phrase Structure Grammar, HPSG, Machine Translation, Synchronous Grammar, Parsing, Natural Language Processing, Computational Linguistics

1. 緒言

本論文は、同期文法に基づく英日自動翻訳の実現に向けて、言語学的に精緻に規定される主辞駆動句構造文法(Head-driven Phrase Structure Grammar, HPSG)[1] のための同期文法(Synchronous Grammar)[2] を提案する。近年、パラレルコーパスからの自動的な学習が可能な統計機械翻訳[3]-[5]が注目を浴びているが、統計機械翻訳は語順の入れ替えや語の置き換えを基本的な翻訳モデルとしているため、統語構造が異なる言語間での翻訳は難しい。同期文法は二言語の文法規則対や辞書項目対からなる文法であり、同期文法を用いて解析することにより、二言語間の統語構造や翻訳が得られる。同期文法を機械翻訳のモデルとすることで、大きく異なる統語構造を持つ言語間においても品質の高い翻訳が実現されることが期待される[6]-[8]。本研究では、英日間の HPSG 同期文法を設計し、その同期文法による機械翻訳の性能評価を行う。

近年、計算機の性能が大きく向上し、また、機械翻訳のための言語的資源の質的量的な増大、および高性能な機械学習の技術が利用可能になったことにより、より洗練された機械翻訳手法を用いることが可能となった。機械翻訳の研究は過去数十年にわたって行われてきたが、90 年代半ば以前は、パラメータを人手により調整するルールによる手法、構文構造を変換する手法、用例に基づく手法および言語学的な文法を用いた意味表現を介する機械翻訳手法を中心に研究がなされてきた。しかし、これらの手法では、ルール作成やパラメータ調整を人手により行うため、ルール間の相互作用による副作用を考慮しながらルールを追加・削除し、パラメータを調整することが難しく、その結果、大規模なシステムを構築することや、性能を向上するための改良が非常に困難であった。言語学的な文法を用いた意味表現を介する機械翻訳はドイツの DFKI の VERB

* 松山市文京町 3 愛媛大学 工学部 情報工学科
Department of Computer Science, Ehime University, Matsuyama, Japan.
原稿受理　平成 23 年 10 月 29 日

Fig.1　A lexical entry pair of nouns in a Synchronous HPSG

MOBIL プロジェクトで過去行われてきたが、実用的なレベルにまでは達していない。これは、実テキストを広く解析できる大規模文法の開発が困難であったこと、および過度に一般化された意味表現を介して言語間のマッピング規則を構築することが困難であったためと考えられる。90 年代にはいって、統計機械翻訳と呼ばれる統計モデルに基づく機械翻訳が IBM の Brown らにより提案され[3]、機械翻訳の領域で注目されている。これは、音声認識や形態素解析でよく用いられている統計的モデルを機械翻訳に応用した機械翻訳方式であり、教師あり・なし機械学習によるルール・パラメータの自動学習が行えるため人手の介在が少なくすみ、上述の副作用を伴わない特長をもつ。しかしこの手法は、単語の翻訳・移動・マッピングを単位とした翻訳手法であり、言語学的制約や句構造を単位とした機械翻訳手法ではないため、日英翻訳のように構文構造が大きく異なる言語間では高精度化が非常に難しい。

ここ数年、統計機械翻訳に句構造をとりこんだ手法[4]や、逆に構文構造を変換する手法や用例に基づく手法に統計モデルを導入する手法[9][10]が研究されており、この一般化として、統計的同期文法[6]-[8]が注目を浴び、研究され始めている。同期文法は、二つの言語を記述する句構造文法間に対応を与えた文法であり、形式的によく定義されたモデルとなっているため統計モデルと相性が良く、文法的な制約により自然な翻訳がなされることが期待されている。同期文法は1960年代AhoとUllmanにより提案されており[2]古くから存在するが、この数年で注目を浴びている理由としては、高速で精度の高い構文解析器が一般に利用可能になったこと、文単位で対応がつけられた 2 言語間の翻訳テキストが利用可能になったこと、および自然言語処理で統計モデルの研究が大きく発展したことが大きな理由と考えられる。しかしながら、LTAG と呼ばれる文法のための同期文法[11] を除く既存の同期文法は、CFG のための同期文法がほとんどであり、言語学的に厳密に定義された文法のための同期文法はまだ提案されていない。言語学的に厳密に定義された文法を用いることで、より文法的な句構造間の対応付けや、より文法的な文の生成が可能になることが期待される。

本研究は、言語学的に精緻に定義された主辞駆動句構造文法(HPSG)のための同期文法をモデル化し、実際に開発することにより、より洗練された機械翻訳を実現することを目的とする。HPSG は言語学的に厳密に定義された語彙化文法であり、そのため、HPSGのための同期文法は、CFGのための同期文法よりも、より文法的な句構造間の対応付けや、より文法的な文の生成が可能となることが期待される。本研究では、同期 HPSG による機械翻訳の実現のため、同期 HPSG の理論化、文法開発、また、BLEU 等のスコアによる評価を行う。

2. HPSG 同期文法

本節では本研究において提案する HPSG 同期文法(以下、同期 HPSG)について説明する。同期 HPSG は、基本的に二言語の HPSG 文法から成る。具体的には、二言語間の対応付けが与えられた辞書項目の対の集合と、二言語間の対応付けが与えられた文法規則の対の集合から構成される。辞書項目の対に文法規則を適用することにより句構造の対が得られ、これを繰り返すことにより同期 HPSG の構文木が得られる。同期 HPSG の構文木は、辞書項目の対や文法規則の対と同様に二言語の構文木の対により表現され、言語間の句構造には対応付けが与えられる。

同期 HPSG の辞書項目は二言語の辞書項目対とそれらの対応からなる。Fig.1 と Fig.2 は日英間における同期 HPSG の辞書項目の例を表している。Fig.1 は“Jack”と“ジャック”の名詞に対する辞書項目の対を表しており、Fig.2 は“eats”と“食べる”の動詞に対する辞書項目の対を表している。PHON:や CAT:は辞書項目の属性を表しており、その右側に書かれている値が各属性の値となる。四角で囲まれた数字は構造共有タグを表しており、

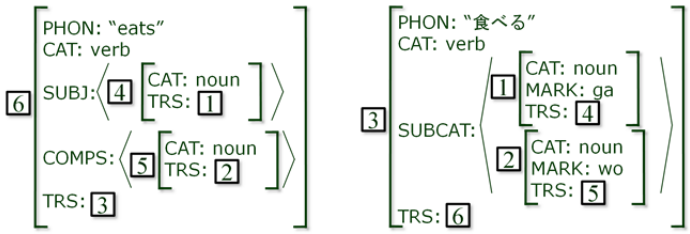

Fig.2　A lexical entry pair of verbs in a synchronous HPSG

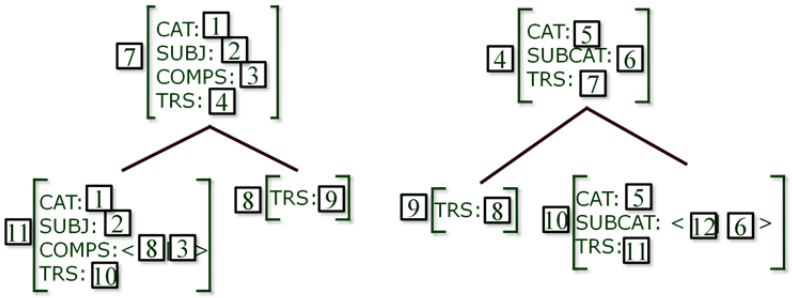

Fig.3　A grammar rule pair for complements in a synchronous HPSG

タグを値として持つ属性を辿ることによりタグに対応する構造を得ることができる。例えば、"Jack"の辞書項目において、TRS:という属性を辿ることにより、対応する"ジャック"の辞書項目が得られる。このように対訳関係にある辞書項目が対となって、同期文法の辞書項目となる。HPSG は語彙化文法と呼ばれ、どのような目的語やどのような主語をとるか、という統語構造を大きく支配する情報は辞書項目に記述される。Fig.2 は動詞に対応する辞書項目であるが、SUBJ:の値はその動詞がとるべき主語を記述し、COMPS:にはその動詞がとるべき目的語のリストを与えている。

同期HPSGの文法規則は、二言語の文法規則対として定義され、文法規則の対応と、二言語間の句や語の線形順序を定義する。Fig.3 は同期HPSGの補語(目的語)をとるための文法規則の例を表している。7,8,11 でタグ付けされた左側の分岐が英語文法における親子関係を表しており、4,9,10 でタグ付けされた右側の分岐は日本語文法における親子関係を表している。それぞれ、7 と 4 が親の構造を表し、11,9 が左の子の構造を表し、8,10 が右の子の構造を表す。それぞれの言語において、部分木と文法規則の子に対応する構造を単一化することにより、親の構造が得られる。Fig.4 は同期 HPSG による構文解析の例を示している。図の左側は"Jack eats bananas"に対する英語の構文木であり、右側はその対訳である"ジャックがバナナを食べる"に対する日本語の構文木を表している。構文解析は、まず、対応する辞書項目の対に対し、文法規則の対を適用し、対の部分構文木(句構造)を得る。図では単語や句構造の対応を点線で表している。同様に得られた部分構文木の対に対し文法規則の対を適用することを繰り返すことにより、全体の構文木の対が得られる。二言語の文の対を構文解析することにより、句構造や単語の対応付けを得ることができる。どちらかの言語の文を解析し、残った言語の文を生成することにより、機械翻訳を実現することができる。

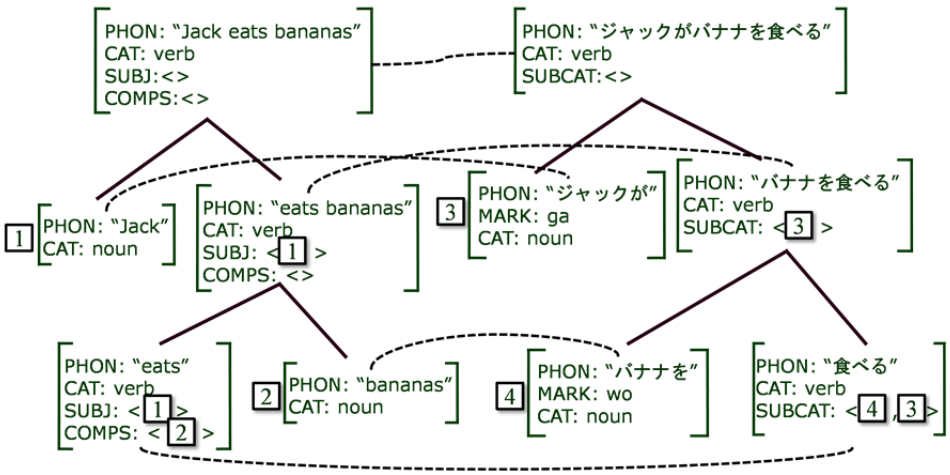

Fig.4 A parse tree pair of a synchronous HPSG

表 1: Moses による性能評価

	NIST	BLEU[13]
wmt08	7.1037	0.2523
wsj	4.2729	0.1193

3. 実験

同期 HPSG を用いて、機械翻訳における性能評価を行う。ただし、本研究においては、同期文法による構文木の構造的対応の性能について評価を行い、単語に対する対訳辞書や単語の翻訳に関する評価は将来の課題とする。同期 HPSG を得るために、日本語、英語の 2 言語の文法が必要となるが、現在、現実世界のテキストを十分に解析できる英語 HPSG 文法[12]は存在するが、日本語 HPSG 文法は存在しない。これは計算機で解析可能な日本語の述語項構造の定義が難しく、既存の木構造付テキストを利用することが難しいことが原因であると考えられる。本研究では、英語文法と独立に日本語文法を作成するのではなく、すでに存在する英語 HPSG[12] を拡張することにより同期 HPSG を実現する。具体的には英語 HPSG の文法規則を拡張し、得られた英語構文木から日本語構文木を自動的に作成する。HPSG は極端に文法規則が少ない語彙化文法であり、十程度の数の文法規則しかもたないため、文法規則の対を記述するのは比較的容易である。同期 HPSG の文法規則には対応する日本語の句構造が記述されているため、英語構文木に対してこの文法規則を適用することにより、対応する日本語の構文木が自動的に得られる。ただし、この手法では単語訳が得られないため、単語アライメント(単語対応付け)のツールとして著名な GIZA++を用いて、対訳文から単語訳を直接得ることにした。

実験では、Penn Treebank の Wall Street Journal に対する 30,117 文の英日対訳テキストを用いた。まず、この英日対訳テキストに対し、統計機械翻訳のツールとしても有名な Moses を用いて、翻訳精度を測定した。25,946 文を学習用コーパス、2182 文をパラメータ調整用コーパス、1989 文をテストコーパスとして実験を行った。日本語の単語分割は MeCab を用いた。表 1 は仏英 WMT08 News Commentary (wmt08, 約 55,000 文)による Moses の性能と、Wall Street Journal (wsj) に対する性能を評価した実験の結果を示す。wsj はスコアが低い傾向にあるが、仏英よりも英日の方が難しい、学習コーパスが小さい、日本語の単語分割が適切でない、wsj の翻訳の質が悪い、などの原因が考えられる。

続いて、同期 HPSG による翻訳性能の評価を行った。ここでは同期 HPSG の句構造対応性能を評価するた

表 2: 同期 HPSG 文法の性能評価

	NIST	BLEU
wsj	6.921	0.2145

め、単語対に関してはテストコーパスからGIZA++によって与えられることとする。表 2 は実験結果を示す。実験結果をみると、得られたスコアが高いことから構造的対応がとれていることがわかる。表 1 の wsj に対する Moses のスコアと比較すると、Moses よりも高いスコアを得ていることがわかるが、これはテストコーパスから単語対を得ているためと考えられる。実際の機械翻訳では、単語訳を自動的に出力する必要があるため、単語訳のモデルを必要とするが、これについては将来の課題とする。また、解析結果を調べると、GIZA++による単語アライメントが失敗している場合が多く、また、構文木の出力に失敗している場合も少なからずあった。これらの問題の解決により、スコアがさらに改善されることが期待される。

4. まとめ

本論文は主辞駆動句構造文法(HPSG)のための同期文法(同期 HPSG)を提案した。同期 HPSG は、辞書項目の対の集合と、文法規則の対の集合からなり、部分構文木も言語間の対として表現される。構文解析は、与えられた部分構文木の対に対し、対の文法規則を繰り返し適用することにより実現される。また、同期 HPSG を用いて、原言語の文に対し構文解析を行い、翻訳先言語の構文木を生成することにより、機械翻訳を実現することができる。

実験的な同期 HPSG を開発し、Wall Street Journal の対訳に対し、機械翻訳の実験を行った。開発した同期 HPSG には単語翻訳のモデルが無いため、単語アライメントツールである GIZA++を用いて、テストコーパスから単語訳を得た。実験結果より、比較的高い翻訳精度が得られたため、同期 HPSG によって良い句構造対応が得られていることがわかった。

同期文法の応用として同期構文解析によるパラレルコーパスの解析が期待されるが、交差などの難しい条件があるためこれについては将来の課題とする。

5. 参考文献

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

[2] A. V. Aho and J. D. Ullman: Syntax directed translations and the pushdown assembler. Journal of Computer and System Sciences, 3:37–56, 1969.

[3] P. E. Brown, V. J. D. Pietra, S. A. D. Pietra and R. L. Mercer: The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 19:263—311, 1993.

[4] P. Koehn, F. J. Och and D. Marcu: Statistical Phrase-Based Translation, In Proc. of HLT-NAACL-2003, pp. 48—54, 2003.

[5] F. J. Och: Minimum Error Rate Training in Statistical Machine Translation, In Proc. of ACL-2003, pp. 160—167, 2003.

[6] David Chiang. A hierarchical phrase-based model for statistical machine translation. In Proc. of ACL-2005, pp. 263–270, 2005.

[7] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics, 23:377–404, 1997.

[8] Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In Proc. of ACL-2001, pp.

523–530, 2001.
[9] 今村 賢治, 大熊 英男, 隅田 英一郎: 句に基づく構文トランスファ方式統計翻訳, 情報処理学会論文誌 48(4), 1809—1819, 2007.
[10] K. Imamura, H. Okuma, T. Watanabe and E. Sumita: Example-based Machine Translation Based on Syntactic Transfer with Statistical Models, In Proc. of COLING-2004, pp. 99—105, 2004.
[11] S. M. Shieber and Y. Schabes: Synchronous Tree-Adjoining Grammars, In Proc. of COLING'90, pp. 253—258, 1990.
[12] Y. Miyao, T. Ninomiya and J. Tsujii: Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In Natural Language Processing - IJCNLP 2004, LNAI3248, pp. 684-693, Springer-Verlag, 2005.
[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu: BLEU: a method for automatic evaluation of machine translation. In Proc. of ACL-2002, pages 311–318, 2002.