構文解析器評価のための GENIA 依存構造コーパス

建石由佳[†]、宮尾祐介[‡]、Kenji Sagae[‡]、大田朋子[‡]、辻井潤一[‡§*]

[†]工学院大学情報学部コンピュータ科学科

[‡]東京大学大学院理工学系研究科情報科学専攻

[§]英国マンチェスター大学　*英国国立テキストマイニングセンター

1. 背景

生命科学関連分野では、論文の急速な増加にともなって、テキストマイニングツールが必要とされるようになっている。そのために自然言語処理技術が応用され、特に、イベント情報を抽出するために述語項構造を抽出する研究が行われている。初期のシステムでは、表層上のパターンマッチングなどの浅い解析を用いていたが、依存関係が階層的であったり、遠く離れる位置におかれた語句同士の依存関係が必要となる場合にはパターンで捕らえることが難しいことも多い。このため、述語項構造の解析手段は深い構文解析へと移ってきている。

構文解析器はいくつかのフォーマリズムに基づくものが提案されており、出力形式等の違いにより、複数の構文解析器の性能を直接比較することが困難である。Caroll ら[1]は、文中の語の依存関係に基づく Grammatical Relations（GR）と呼ぶ構造を用いて、異なるフォーマリズム間の出力形式の差異を吸収し構文解析器の精度を比較する方法を提案した。GR は、Penn Treebank (PTB)形式のフォーマリズムに基づく構文解析器と LFG や HPSG などの語彙化文法に基づく構文解析器との精度の比較に用いられている[2-6]。新聞テキストについては Penn Treebank[7]の Wall Street Journal コーパスの一部に GR 構造を人手で付与したコーパス[5,8]（以下 WSJ-GR と呼ぶ）を使うことができるので、フォーマリズムによらない精度評価の枠組みが提供されているといえる。

一方、生命科学論文に対する構文解析器の評価では、過去、Stanford Dependency Scheme[9]に基づくコーパスが用いられてきた[10,11]。これには、Penn Treebank形式からStanford Dependency形式への自動変換スクリプトが供給されている[1]ことが背景にあり、過去の研究[10,11]でもTreebank形式のコーパスから自動変換したコーパスが用いられている。ところが、自動変換スクリプトは正確ではなく、また、どのようなバグがスクリプトに存在するのかに関する記述が与えられていない。このため、自動変換によって生成されたコーパスはGold Standardであるか、とくにPTB形式とそれ以外の形式の構文解析器の比較で正確な評価ができているかどうかが疑わしい。

2. コーパスの概要

前節に述べたことから、われわれは新たに生命科学分野での構文解析器の精度を比較する目的でその Gold Standard となるコーパスを作成することにした。形式は Carroll らの GR 形式を採用する。これは、GR が過去、生命科学分野以外で、複数のグループにより構文解析器の評価に用いられてきたこと、新聞分野で人手でタグ付けされた Gold Standard コーパスが存在するため、他分野のテキストでの精度と生命科学分野テキストでの精度を比較するためにも都合がよいことが主な理由である。

テキスト

コーパスのベースとなるテキストはGENIAコーパス[12]から以下の基準を満たすものを選んだ。

１）MeSH[2]キーワード「NF kappaB」を持つ

[1] http://nlp.stanford.edu/software/lex-parser.shtml

[2]論文アブストラクトデータベース PubMed においてキーワードとして用いられる統制語彙

2）PubMed Central[3]に登録されている
3）原らの実験[13]において訓練データに使用されていない

基準 1)を満たすアブストラクトは小田ら[14]が転写因子 NF κ B のパスウェイを作成するのに用いたセットで、パスウェイ情報抽出において手がかりとなる表現に対するマークアップが存在する。基準 2)は、本文テキストデータの入手が容易であるということを意味する。基準 2)を満たすアブストラクトを採用することによって、将来、情報抽出の対象がアブストラクトから本文に移った場合にコーパスの拡張を容易にすることを期待している。基準 3)は、原らの実験に用いられた構文解析器も評価の対象であることから、公平性のために設けた。GENIA コーパス中には上の 3 基準をすべて満たすアブストラクトが 50 あり、文の総数は 486 であった。コーパスのサイズとしては大きいものではないが、WSJ-GR のサイズ（560 文）との比較から評価用には十分な大きさであると判断してこの 50 アブストラクトを用いた。

スキーマ

GR スキーマ[8]では、コーパス中の各文に対し、*rasp* という構造が付与される。これは、文中の語の依存関係を、head-dependent のペアをその依存関係の種類とともに並べ挙げることにより表現するものである。われわれは、この *rasp* 構造を XML 化し、さらに *named_entities* 要素を付与した。要素 *named_entities* はも文ごとに付与され、子要素として *term* を持つ。１つの *term* 要素は GENIA 専門用語コーパスでマークアップされている用語のうち複数語からなるものであり、*id* 属性と *sem* 属性を専門用語コーパスから引き継いでいる。さらに、文中の位置を示す *span* 属性を持つ。複数語からなる用語は rasp 要素中ではその id のみが参照される。句構造（構文木）についてはこのコーパスではマークアップせず、将来 GENIA Treebank[15]の構造を取り込む予定である。また、作業中の疑問点をメモするために使う *note* 要素(内容は自由記述のコメント)を含むことがある。タグ付けの例を図１に示す。図では見易さのために WSJ-GR と同様の形式にし、*term* の sem 属性は省略している。図中の*T1*などは *term* に対する id 参照である。

```
sentence( id(96099434.1)
named_entities(
term( id(T1) span(1:3)(Protein kinase C-zeta))
term( id(T2) span(5:7)(NF-kappa B activation))
term( id(T4) span(9:12)(human immunodeficiency
             virus-infected monocytes))
sentence_form(Protein kinase C-zeta mediates
NF-kappa B activation in human immunodeficiency
virus-infected monocytes . )
rasp(
(ncsubj mediates:4 *T1* _)
(iobj mediates:4 in:8)
(dobj mediates:4 *T2*)
(dobj in:8 *T4*)
))
```

図1:タグ付け例

3. タグ付けの実際

タグ付けは RASP システム[16]の出力を人手で修正することによって行った。コーパス作成の主な目的が Penn Treebank 形式の出力を出す解析器とそれ以外の形式の出力を出す解析器の精度比較であるため、公平性の観点から、Penn Treebank に準じる GENIA Treebank とは独立に作業を行っている。作業は一人の作業者（著者の一人）が行い、不明な点は WSJ-GR を参照しそれに準じた。作業を行うちちに GENIA コーパスに特有と思われるいくつかの問題点が出てきたが、それらに対しては暫定的なルールを定めた。たとえば、カンマなどの記号を用いない同格表現（nuclear factor NF kappa B など）については[8]にタグ付けの指針がない。このケースについては、修飾-日修飾の関係を表す *ncmod* 関係の下位分類 *ta* (text adjunct : [8]にはこの使用法に関する指針がない)を流用することで対処した。

[3] 米国 National Institutes of Health (NIH) による論文アーカイブで、医学・生物学関連の 200 以上の雑誌からの論文が自由にダウンロード可能である

現在、50 アブストラクトに対する初期作業を終え、修正作業に入っている。修正に先立ち、初期作業者により疑問点として *note* タグが付けられたところを分析した。主なものを分類すると次のようになった。

1）動詞の後の前置詞句を格要素（*iobj*）扱いにするのか修飾句(*ncmod*)扱いにするのか

2）省略の絡む並列句の構造をどう付けるか

3）前置詞句の係り先

4）専門用語とされている語句の一部に依存する関係を付けたい

5）トークンの一部に依存する関係を付けたい

1）については、[8]にも明確な基準が提示されておらず、WSJ-GR でも動作性の名詞に対する意味上の動作主体が of により示される場合では多くが *iobj* 扱いである一方受身の主語を示す by 句が統一的に *ncmod* 扱いである、など、わかりやすいポリシーが見えない。これについては個別の動詞および動作性名詞ごとに基準を定める必要があろう。

2）は実は二つの問題を内包している。一つは実際に何が並列化されているかの判断が難しいことである。たとえば、「CD4(+) and CD8(+) T lymphocytes」は「CD4 と CD8 の両方が発現した T 細胞」（Collective な読み）をさすのか「CD4 が発現した T 細胞と CD8 が発現した T 細胞」（Distributive な読み）をさすのかが判断しづらい。また、類似の例として並列句の直近に修飾句がある場合（「normal T-cell activation and growth」など）修飾の対象が並列される句の両方なのか片方なのかがわかりにくいこともある。このような例については個々のケースごとに解決していくしかないが、ここで、もう一つの問題が発生する。その第 2 の問題は Distributive な読みについては GR のスキーマで表現する手段がないことである。GR では文中の省略を示すためには *ellip* というマーカーを置く以外の表現方法がなく、また、*ellip* と省略された（もとの）要素との間の対応を付ける依存関係が定義されていない。このため、GENIA Treebank においては上の例に挙げたような例について NULL 要素の id 参照と種別によって Distributive かどうかを表現することが可能であったが、現在の GR スキーマではその区別を落とさざるを得ない。省略の絡む並列は論文アブストラクトには多数存在し、また、何が並列されているかを知ることは情報抽出の観点から必要であると思われるため、これらを区別する仕組みを GR に加える必要があろう。

3）は 2）（の前半）と同様に意味的な区別が難しいことに起因し、個別に専門家の判断により解消する。

4）に関して、GENIA 専門用語をあたかも 1 語であるかのように扱う現在のアプローチは、物質名など内部構造を問題にしても意味のない句の分析を省略できるという意味でアノテーション作業を簡略化し、また、GENIA Treebank や BioInfer[17]など他のコーパスでとられている「並列の絡まない名詞句の内部は分析を省略する」というポリシーにもほぼ合うものになる。しかし、GENIA コーパスにおける専門用語の概念は「名前」よりも広く、しばしば細胞の発達段階、場所、生物種などの修飾語を含んでいる。この修飾語をさらに別の語が修飾する場合に 4）のような問題が起こる。たとえば「more mature cell lines」という句では明らかに more が mature を修飾するのに、「mature cell lines」が専門用語と認定されている。一方、「simian immunodeficiency virus」のように同じような構造をしていても名前として一般に認識されている語句もあり、区別が難しい。将来は専門用語の内部も解析し、どうしても解析できない場合のみ term として残す、という方針に切り替えたほうがよいように思われる。その際、現在のコーパスに付けられている term を集め、それをまとめて分析した上で Gazetteer として利用することで、専門用語内部に統一的な分析ができると期待できる。

5）は「N- and C-terminal」のようなケースで、GENIA Treebank が Penn Treebank の基準をそのまま採用して、ハイフン（ここでは問題になっていないがスラッシュも）を単語境界と認めていないことに起因している。これについては GENIA POS コーパスおよび GENIA Treebank の単語境界の基準を見直すことを検討している。

4. まとめ

GENIA コーパスの一部に依存構造を付与したコーパスを作成した。現在、初期作業が終了し。修正作業とともにスキーマの追加・修正、基準の明確化を行っている段階である。作成したコーパスは GENIA コーパスの一部として公開する予定である。

参考文献

[1] J. Carroll, E. Briscoe, and A. Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. Proc. LREC 1998.

[2] J. Preiss. 2003. Using grammatical relations to compare parsers. Proc. EACL 03.

[3] R. Kaplan, S. Riezler, T. King, J. Maxwell, A. Vasserman, and R. Crouch. 2004. Speed and accuracy in shallow and deep stochastic parsing. Proc. HLT/NACL 2004.

[4] E. Briscoe and J. Carroll. 2006. Evaluating the Accuracy of an Unlexicalized Statistical Parser on the PARC DepBank. Proc. COLING/ACL 2006, Sydney, Australia.

[5] S. Clark and J. Curran. 2007. Formalism-Independent Parser Evaluation with CCG and DepBank. Proc. ACL 2007, Prague, Czech Republic.

[6] Y. Miyao, K. Sagae and J. Tsujii. 2007. Towards Framework-Independent Evaluation of Deep Linguistic Parsers. Proc. Grammar Engineering across Frameworks.

[7] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. Computational Linguistics, 19:313-330.

[8] E. Briscoe. An introduction to tag sequence grammars and the RASP system parser, Technical Report (UCAM-CL-TR-662), Cambridge University Computer Laboratory, 2006.

[9] M-C. de Marneffe, B. MacCartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. Proc. LREC 2006.

[10] S. Pyysalo, F. Ginter, V. Laippala, K. Haverinen, J. Heimonen, and T. Salakoski. 2007. On the unification of syntactic annotations under the Stanford dependency scheme: A case study on BioInfer and GENIA. Proc. BioNLP Workshop at ACL 2007.

[11] A. B. Clegg and A. J. Shepherd. 2007. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC Bioinformatics 8:24.

[12] J-D. Kim, T. Ohta, Y. Teteisi and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1).

[13] T. Hara, Y. Miyao and J. Tsujii. 2007. Evaluating Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser. Proc. IWPT 2007.

[14] K. Oda, J-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii. 2007. New challenges for text mining: Mapping between text and manually curated pathways. Proc. LBM 2007.

[15] T. Ohta, Y. Tateisi, J-D Kim, A. Yakushiji and J Tsujii. Linguistic and Biological Annotations of Biological Interaction Events. Proc. LREC 2006.

[16] E. Briscoe, J. Carroll and R. Watson. 2006. The Second Release of the RASP System. Proc. COLING/ACL 2006 Interactive Presentation Sessions.

[17] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen and T. Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics 8:50