深い構文解析を用いた関係抽出のための文簡易化

三輪 誠　　辻井 潤一
東京大学　東京大学/マンチェスター大学/NaCTeM

概要

1 文中の 2 つの固有名間の関係抽出においては，その固有名を含む構文解析結果の一部を元にカーネルや素性ベクトルを設計し，分類を行う手法が広く利用されている．この部分構造からは関係抽出に不必要な情報を取り除く必要がある．このために，我々は，蛋白質間相互作用抽出を対象に，関係抽出のための文簡易化手法について提案する．深い構文解析の結果を利用し，対象とする固有名 (蛋白質) に着目することで，単純かつ一般的な，また，関係抽出に必要な情報をできるだけ失わないルールを 12 個設計した．この単純なルールを用いた文の簡易化により，用いた比較的小さなコーパスについて，関係抽出結果を改善できた．

1　はじめに

日々増え続ける文章から固有名 (entity) の間の関係を人手で抽出するのは手間が掛かり，自動的な関係の抽出が必要とされている．この抽出の中でも，1 文中に現れる 2 つの固有名間の意味的な関係を判断するタスクは最も基本的であるため広く研究されている．このような関係抽出タスクにおいては『人名と企業』などの一般的な関係を含んだ ACE RDC (Automatic Content Extraction Relation Detection and Characterization) [*1] 2003, 2004 コーパスや『蛋白質間相互作用抽出』という生物医学における関係を含んだ AIMed コーパス[*2]が良く用いられる．この関係抽出タスクは，2 つの固有名間に関係があるかないか，という分類問題として扱われている．

近年，構文解析結果を基にした機械学習手法が関係抽出において有効であると分かってきた．これらの手法は固有名のペアを含む構文解析結果の部分構造を選択 (もしくは区別) したカーネルや素性ベクトルを設計し，利用する．この部分構造には固有名間の最短経路 [1] や固有名のペアを含む部分木 [2] などが用いられる．これらの部分構造には同格など関係の抽出には不必要な情報が含まれている．

本稿では文簡易化により 2 つの固有名間の関係抽出に不必要な情報を文から取り除く手法について提案する．関係抽出タスクとしては蛋白質間相互作用抽出を対象にする．この文簡易化により，2 つの固有名間の関係の表現を簡易なものにし，既存の関係抽出手法がより有効に働くことが期待できる．対象とする固有名に着目することで関係に関する表現を失うリスクを避けながら，構文解析結果の利用を前提とすることで単純かつ一般的な 12 個のルールを設計した．図 1 のように，このルールを繰り返し適用することで，関係抽出のための文簡易化を行った．本稿の貢献は，構文解析結果を利用した容易かつ一般的な文簡易化，対象とする固有名に着目するという設計指針に基づいたルールの提示，文簡易化による (小さなコーパスにおける) 関係抽出の性能向上，である．

2　関連研究

関係抽出において，構文解析結果から固有名間の関係により有効な部分構造を得る手法として，いくつかの木カーネル [2, 3] が提案されている．また，部分構造に含まれなかった情報を補う手法として，グラフカーネル [4] や合成カーネル [2, 3, 5] を用いた手法がある．これらのシステムの中で我々は評

[*1] http://www.ldc.upenn.edu/Projects/ACE/
[*2] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/

2) IGF-1, IGF-2, and insulin also induced a Ca2+ mobilization from the endoplasmic reticulum: *phospholipase C* (PLC) inhibitors, neomycin, or U-73122 partially blocked the intracellular [Ca2+]i increase induced by IGF-1 and *insulin* and totally inhibited the effect of IGF-2 .
⇒ *phospholipase C* (PLC) inhibitors, neomycin, or U-73122 partially blocked the intracellular [Ca2+]i increase induced by IGF-1 and *insulin* . **[文選択]**
⇒ *phospholipase C* (PLC) inhibitors, or U-73122 partially blocked the intracellular [Ca2+]i increase induced by IGF-1 and *insulin* . **[等位]**
⇒ *phospholipase C* (PLC) inhibitors partially blocked the intracellular [Ca2+]i increase induced by IGF-1 and *insulin* . **[等位]**
⇒ *phospholipase C* (PLC) inhibitors partially blocked the intracellular [Ca2+]i increase induced by *insulin* . **[等位]**
⇒ *Phospholipase C* inhibitors partially blocked the intracellular [Ca2+]i increase induced by *insulin* . **[括弧]**

図 1　文簡易化の例．IEPA コーパス．(斜体は対象とする蛋白質．)

価に用いた AkaneRE[*3]を紹介する．AkaneRE は bag-of-words，最短経路，グラフカーネルを利用しており，蛋白質間相互作用抽出において最高の精度を挙げているシステムの 1 つである [5].

関係抽出における文簡易化としては, Link Grammar による構文解析器の精度向上を目指した bioSimplify [6] を用いた手法がある．この手法では AIMed コーパス [7] において，適合率をあまり減らすことなく，再現率の向上に成功している．

3　提案手法

提案手法では，構文解析器 Mogura[*4] [8] の出力に適用可能な単純なルールを用いて文を簡易化する．Mogura とは HPSG 文法を用いた深い構文解析を行い，構文構造および述語項構造を出力する構文解析器である．我々の手法は，bioSimplify と違い，構文解析結果を利用し，対象とする固有名に着目した簡易化を行う．

3.1　ルール

我々は次のような 2 種類 12 個のルールを設計した．(括弧はルールの数)

- 文選択ルール
  重文・複文　固有名の含まれる文を選択．(1)
- 固有名置換ルール
  同格　固有名が固有名を含まない他の句と同格関係にある場合，同格関係を含む句と固有名を置換．(2)
  例示　(such as, including について) 固有名が例示する，もしくは例示される句にある場合，例示を含む句と固有名を置換．(4)
  括弧　固有名が括弧の (述語項の) 項 (= 括弧の前か中) にある場合，括弧の項を含む句と固有名を置換．(2)
  等位　1 つの固有名が等位接続詞 (and, or など) の項である場合，等位接続詞の項を含む句と固有名を置換．(3)

ルールの例として同格ルールの 1 つを図 2 に示す．同格ルールでは，“,” の述語が “appos_arg12” (同格) であり，その項である “protein” と “A” が同格関係であるという構文解析結果に基づき，その同格の含まれる句を “A” に置換する．

3.2　文簡易化

文簡易化の疑似コードを図 3 に示す．文を構文解析した結果に対し，5-12 行目でルールを 1 つずつ適用し，適用できたら簡易化された構文解析結果から文を再構築する (8-10 行目) . これを繰り返し，どのルールも適用できなくなったら，最後に再構築された文を出力する．(14 行目)

4　評価

評価として，ルール適用による蛋白質間相互作用抽出の性能の違いを示す．評価には蛋白質間相互作用抽出を対象とした 5 つのコーパス (AIMed・

[*3] http://www-tsujii.is.s.u-tokyo.ac.jp/~satre/akane/
[*4] http://www-tsujii.is.s.u-tokyo.ac.jp/enju/

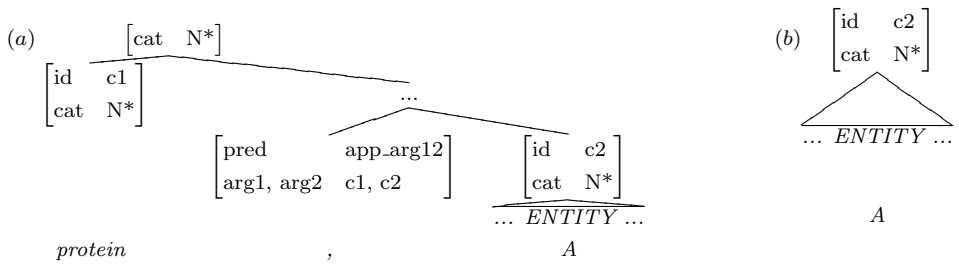

図 2　同格ルール．構文解析結果の (a) を (b) に置換．

表 1　各コーパスのルール適用による AUC-ROC (Area Under the Receiver Operating Characteristic Curve) の変化とルール適用回数．(太字はルール適用により改善したもの．)

		AUC-ROC				適用回数		
コーパス	例数	適用前	文選択	固有名置換	全ルール	文選択	固有名置換	全ルール
BioInfer	9,653	84.2	**84.5**	81.7	81.9	3,819	25,003	27,555
AIMed	5,834	85.5	**85.6**	84.6	84.1	2,314	9,734	11,556
IEPA	817	85.1	**85.4**	**87.8**	**88.2**	289	1,258	1,474
HPRD50	433	84.5	83.6	**87.2**	**86.2**	145	695	829
LLL	330	84.4	**84.6**	**91.2**	**92.2**	135	622	718

```
1: S ← input sentence
2: repeat
3:    reset rules
4:    P ← parse S
5:    repeat
6:       r ← next rule {null if no more rules}
7:       if r is applicable to P then
8:          P ← apply r to P
9:          S ← sentence extracted from P
10:         break
11:      end if
12:   until r is null
13: until r is null
14: return S
```

図 3　文簡易化の擬似コード．

BioInfer・HPRD50・IEPA・LLL) [*5]を用いた．蛋白質間相互作用抽出には AkaneRE を用いた．AkaneRE の学習器である SVM のパラメータ C は 1 に固定し，一般化された蛋白質間相互作用抽出のために蛋白質名は隠し (2 つの対象とする蛋白質を P1, P2，それ以外は PROT) アブストラクト単位での 10 分割交差検定で評価した．ここで，ルールの適用順については，事前実験において影響が見られなかったため，重文・複文 → 同格 → 例示 → 括弧 → 等位の順に固定した．

それぞれ 2 グループのルールを適用した結果を表 1 に示した．表 1 より，大きなコーパス (BioInfer, AIMed) では文選択ルール以外は改善がないが，小さなコーパス (IEPA, HPRD50, LLL) においては，HPRD50 における文選択ルールの適用を除き，ルールを利用することで改善が見られていることがわかる．また，ルールの適用回数については，全てのコーパスについて，1 文につき平均 2 回程度適用されていることがわかる．

エラー分析として，IEPA コーパスに対する結果の解析を行った．IEPA コーパスにおいて文の簡易化を行った結果の例を図 1 に挙げた．これは文簡易化により新たに見つかるようになった 46 例の正例の 1 つである．最初の文選択において，最後の等

[*5] http://mars.cs.utu.fi/PPICorpora/

Overall, our data indicate that the normal processing of *PrP(c)* is up-regulated by protein kinase C but not *protein kinase A* in human cells and murine neurons.
⇒The normal processing of *PrP(c)* is up-regulated by *protein kinase A* in human cells and murine neurons.

図 4　文簡易化に失敗した例．IEPA コーパス．(斜体は対象とする蛋白質．)

位接続詞以降がなくなっているのは構文解析の失敗のためであるが，最後に簡易化された例は，元の文と同じく 2 つの蛋白質が相互作用している記述となっている．この例では相互作用に関係のない記述をルールの繰り返しの適用により排除できたことがわかる．一方で文簡易化により見つからなくなった正例は 4 例あり，その内の 3 つが括弧ルールにより括弧の中にあった略称がなくなっていることが原因であった．この略称は一般化された蛋白質相互作用抽出のためには利用しない方が良い情報であり，これが今の特徴空間で発見できないことが分かったのは，AkaneRE の改善のためには良い結果であると考えられる．また，文簡易化の失敗した例を図 4 に示した．これは元の文でも AkaneRE が発見できなかった例であり，AkaneRE の改善が必要なものである．この文では否定 (not) が文簡易化の過程で消えてしまっているため，相互作用していない蛋白質が文簡易化により相互作用しているような記述になっている．

5　おわりに

本稿では，構文解析結果に基づいたルールを利用した関係抽出のための文簡易化手法について提案した．構文解析器による一般化を利用した 12 個の少ないルールを繰り返し利用することで，多くの文を簡易化することができた．また，利用した中でも小さなコーパスについては，この簡易化により関係抽出の性能を大きく改善することができた．

今後の課題としては，今回改善が見られなかったコーパスについて，その原因の分析が挙げられる．また，蛋白質間相互作用抽出以外の関係抽出におけるルールの利用可能性についても評価したい．

参考文献

[1] Razvan C. Bunescu and Raymond J. Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of HLT '05*, pages 724–731, 2005.

[2] Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. A composite kernel to extract relations between entities with both flat and structured features. In *Proceedings of ACL-44*, pages 825–832. Association for Computational Linguistics, 2006.

[3] Truc-Vien T. Nguyen, Alessandro Moschitti, and Giuseppe Riccardi. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of EMNLP 2009*, pages 1378–1387. Association for Computational Linguistics, August 2009.

[4] Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. A graph kernel for protein-protein interaction extraction. In *Proceedings of the BioNLP 2008 workshop*, 2008.

[5] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, June 2009.

[6] Siddhartha Jonnalagadda and Graciela Gonzalez. Sentence simplification aids protein-protein interaction extraction. In *Proceedings of LBM 2009*, pages 109–114, November 2009.

[7] Razvan C. Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, 2005.

[8] Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Efficient hpsg parsing with supertagging and cfg-filtering. In *Proceedings of IJCAI'07*, pages 1671–1676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.