メタ関係を利用したテキストからの人体部位関係の抽出

荒牧英治† 今井健† 梶野正幸‡ 美代賢吾† 大江和彦†
†東京大学医学部附属病院 ‡臨床医学オントロジー研究会

{aramaki, ken, kohe}@hcc.h.u-tokyo.ac.jp
kajino@medical-ontology.jp, miyo-tky@h.u-tokyo.ac.jp

1 はじめに

医療分野では，これまで様々なリソースによって疾患名，人体の部位名などの用語が整理されたきた．例えば，代表的な医療用語データベースである SNOMED-CT[6] には，人体各部の部位名に約 3 万項目がさかれている．しかし，部位同士の関係については，同義関係や部分全体関係など少数の基本的な関係のみしか扱われてこなかった．もし，より詳細な部位の関係，例えば，部位同士の隣接関係などが分かれば，肋骨の損傷が，隣接する肺に影響を及ぼす可能性があるなどの，高度な推論を行う際の知識ベースとして利用でき，有益であろう．しかし，多くの部位同士の関係をすべて人手で記述するのは，大変な労力であり，我々はテキストからの自動的な獲得を目指している．

このタスクは，関係抽出の一種であるといえる．関係抽出の際にはタグ付きコーパスからの教師あり学習がしばしば用いられる．しかし，臨床医療分野のタグ付きコーパスは存在しないので，我々は，最初に「心臓 PART-OF 胸部」などシードとなる関係を与え，それを利用して新たな関係を発見していくブートストラップ手法 [1] を用いる．

ここで，問題となるのは，ブートストラップを繰り返すにつれノイズが入り，誤りが拡大する可能性である．

そこで，本研究では，求めたい関係同士が制約をもつことに着目する．例えば，部分全体関係（PART-OF）と同じサイズである関係（SAMESIZE-AS）という 2 つの種類の関係（以降，関係タイプと呼ぶ）があるとする．この場合，次の 2 つの関係は両立しない．

心臓 PART-OF 胸部
心臓 SAMESIZE-AS 胸部

これは，両者が同じサイズであるならば，一方が他方に含まれることは考えられないからである．よって，この場合，どちらかの関係が誤っていると考えられる．

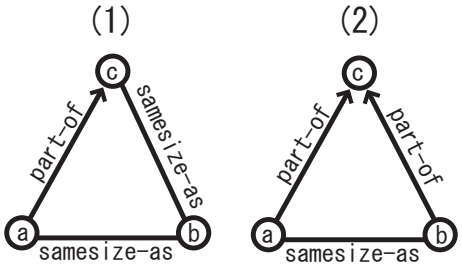

図 1: メタ関係の例

先の例は，単純な例であったが，図 1 の (1) のように複数の関係をたどって判明する矛盾もある．この例では b と同サイズであるはずの 2 つの部位（a と c）が部分全体関係となり矛盾が生じている．本稿では，このような関係タイプ間の関係を**メタ関係**と呼び，メタ関係に矛盾があった場合は関係を棄却することを考える．

次に問題となるのは，メタ関係の矛盾をどのように定義するかである．先のように，SAMESIZE-AS や PART-OF という 2 つの関係タイプしか扱わないのなら人手で記述することも可能であろう．しかし，関係の種類がより多くなると，調整／変更が困難であり，また，汎用性に乏しいという欠点がある．

そこで，本稿では判明している関係の中で，関係タイプがどのように共起するか調べ，頻出するものほど妥当だと考える．例えば，先の図 1(1) は，既知の関係の中では出現しないと思われる．一方，(2) のような矛盾を含まないメタ関係は，数多く出現するであろう．これに着目して，提案手法はメタ関係の頻度を数え，頻度が閾値を下回った場合は，矛盾するとみなす．

実験の結果，提案手法の妥当性を確かめることができたので報告する．

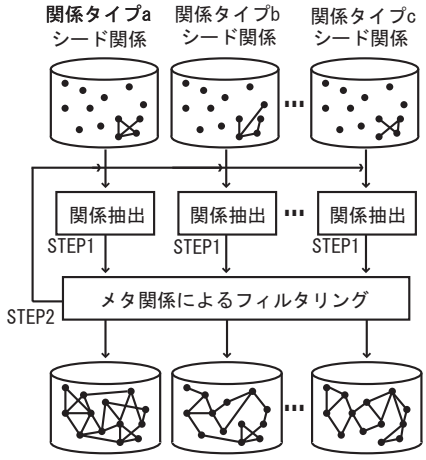

図 2: 処理の流れ

2 提案手法

まず，タスクを説明する．本タスクは，(1) 部位用語のリストと (2) 関係タイプごとシード関係，(3) 大規模コーパスが与えられた際に，各関係タイプごとに (1) 間の関係を可能な限り発見する．

手法の流れは，ブートストラップによる関係抽出を行う処理 (STEP1) と，メタ関係を用いてフィルタリングを行う処理 (STEP2) の 2 つに大きく分かれ，これらを交互に繰り返す．STEP1 は関係タイプごとに独立に行う．STEP2 は共通の処理である．図 2 に処理の流れを図示する．

STEP1:関係抽出

関係抽出にはブートストラップ手法 [1] を用いる．これは，最初に少しだけの教師（シード）を使う手法である．例えば,「**下顎骨** PART-OF **顎顔面骨**」などをシードにして，PART-OF 関係を示す強い証拠となりそうなコンテキストを見つける (STEP1-1)．そして，コンテキストが見つかったら、今度はそのコンテキストに現れるほかの関係を抽出する (STEP 1-2)．

STEP1-1: 関係抽出パターンの生成

最初に，現在判明している関係それぞれについて，関係に含まれる用語が共起する文をコーパスから収集する．例えば,「**下顎骨** PART-OF **顎顔面骨**」が分かっている場合は，“**下顎骨**” と “**顎顔面骨**” が共起する文を集める．

次に，その文を形態素解析器 [5] により形態素解析し，前後 n 語のウィンドウ*を用いて関係抽出パターンを生成する．例えば,「特に【顎顔面骨】のなかでも【下顎骨】など咬合や咀嚼機能に...」という文で出現するならば，得られる関係抽出パターンは，以下のようになる：

【a】|の|なか|でも|【b】
【a】|の|なか|でも|【b】|など
【a】|の|なか|でも|【b】|など|咬合
⋮
【a】|の|なか|でも|【b】|など|咬合|や|咀嚼|機能
⋮
特に|【a】|の|なか|でも|【b】
⋮
特に|【a】|の|なか|でも|【b】|など|咬合|や|咀嚼|機能

.

次に，関係抽出パターンの確信度を調べる．これは得られた関係抽出パターンが，どのくらい現在判明している関係とマッチしているかを示す次式によって行う：

$$\text{パターン確信度 } (p) = \text{conf}(p)\log(countMatchedRel(p) + C_1),$$

$$\text{conf}(p) = \frac{countMatchedRel(p)}{countRel(p)},$$

ただし，$countRel$ はパターン p によって得られる関係の数，$countMatchedRel$ はパターン p を用いて得られる関係のうち既知の関係とマッチするものの数,C_1 は定数である．

パターン確信度が閾値 Th よりも高ければ，このパターンを採用する．

STEP1-2: 関係候補の抽出

採用されたパターンを用いて，新たな関係候補を抽出する．この処理は，各関係タイプごとに行う．また，すべての関係タイプについて，STEP1 の処理が終わると STEP2 へ進む.

STEP2:メタ関係によるフィルタリング

STEP2 では，STEP1 で発見された関係候補について，まず，(1) その周辺に出現する関係タイプの頻度と，(2) 対応候補と (1) の両方が共起する頻度を調べ，(2)/(1) が十分高ければ妥当なメタ関係とみなし，対応候補を採用する．本稿では，(1) を $ContextSet$ と呼び，(2) を $TargetSet$ と呼ぶことにし，次節の通り定義する．

*後述する実験では $n = 5$ とした.

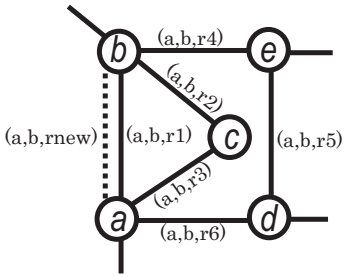

図 3: 関係候補とその周辺の関係

$ContextSet$ と $TargetSet$ の定義

まず，説明を簡便にするため，ノーテションを述べる．用語 a と用語 b の間に関係 r があった場合†，次ののように表記する：(a, b, r).

周辺は，関係候補と接続している n 本の関係のループを使って定義する．例えば，図 3 の例では，$n = 2$ の場合，次の関係ループが存在する: $(a, b, r_{new}) \rightarrow (b, c, r_2) \rightarrow (c, a, r_3)$．ここで，このループから，用語の情報を汎化したものを $TargetSet(n)$ とする: 例えば，図 3 の例では次のようになる:

$$TargetSet(2) = \{(*, *, r_{new}) \rightarrow (*, *, r_2) \rightarrow (*, *, r_3)\},$$

ただし，*はワイルドーカード（どのような用語も当てはまるもの）と考える．

また，$TargetSet(n)$ からさらに関係候補の情報を汎化したものを $ContextSet(n)$ とする．例えば，図 3 の例では次のようになる:

$$ContextSet(2) = \{(*, *, *) \rightarrow (*, *, r_2) \rightarrow (*, *, r_3)\}.$$

メタ関係の妥当性

先に定義した $TargetSet(n)$ と $ContextSet(n)$ を用いて，関係候補のメタ関係の妥当性を次のように定義する：

$$\text{メタ関係の妥当性} = \sum_{n=1}^{N} w_n \cdot \text{conf}(n) \cdot \log(TargetSet(n) + C_2),$$

$$\text{conf}(n) = \frac{\#\text{of}TargetSet(n)}{\#\text{of}ContextSet(n)},$$

ただし，w_n は n ごとの重みである．N は周辺の関係をどこまで見るかを示す定数‡,C_2 は定数である．

このメタ関係の妥当性が高いものから順に C_3 個の関係候補を採用し，STEP1 に戻る．新しく発見される関係がなくなるまでこれを繰り返す.

†r は関係タイプと方向の 2 つの情報を持つものとする.
‡後述する実験では，N=3 とした.

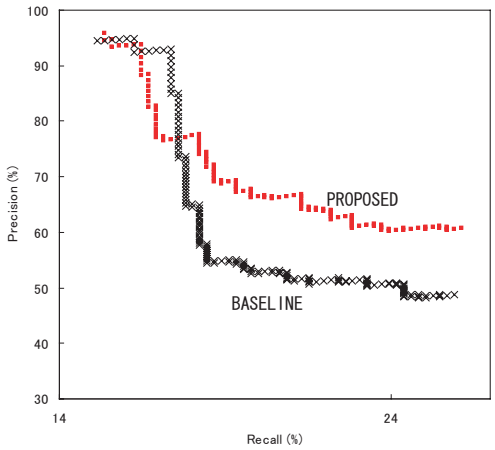

図 4: PROPOSED と BASELINE の精度

3 実験

3.1 実験設定

提案手法の妥当性を確かめるため，実験を行った．これは，次の 2 つの手法を比べることによって行った.

1. PROPOSED: 提案手法.

2. BASELINE: メタ関係によるフィルタリングを行わない場合（1 回のループごとにパターン確信度の高い関係 C_3 個を出力する）.

コーパスは神戸大学解剖学講義ノート§，慶応大学電子教科書¶から抽出した文 (50,000 文) を用いた．部位名は標準病名マスター [3] の修飾語リストに掲載されている人体部位用語のうちコーパスに 5 回以上出現する部位名 427 語を扱った．部位間の関係タイプとそのシード関係の数を表 1 に示す．また，シードはすべて人体頭部の部位の関係を用いた．定数は，$C_1 = 1$,$C_2 = 1$,$C_3 = 10$,$Th = .1$ とした.

評価は，PART-OF 関係に対して，正解となる対応を人手で記述して (746 関係)，次の適合率と再現率を用いて行った:

$$\text{適合率} = \frac{\#of(\text{出力された関係} \cap \text{正解関係})}{\#of \text{ 出力された関係}},$$

$$\text{再現率} = \frac{\#of(\text{出力された関係} \cap \text{正解関係})}{\#of \text{ 正解関係}}.$$

§http://www.lib.kobe-u.ac.jp/products/anatomy/
¶http://web.sc.itc.keio.ac.jp/ funatoka/anatomy.html

表 1: シード関係の数

関係タイプ	シード数
PART-OF (部分全体関係)	182
ADJACENT-TO* (隣接関係)	50
SAMESIZE-AS* (同サイズの関係)	30
SAMEGRANULARITY-AS* (同粒度の概念関係)	480

*印のついた関係タイプは無向の関係とし，$a \rightarrow b$ と $b \rightarrow a$ を区別しないものとする．

3.2 結果

精度は，図 4 のとおりである．PROPOSED の精度が大部分において，BASELINE を上回っているため，提案手法の枠組みは機能していると考えられる．

表 2 に，PROPOSED によって得られた関係と棄却した関係の例を示す．

3.3 考察

実験結果では，PROPOSED の方が精度が高いとはいえ，両手法とも，繰り返し回数が進むにつれ適合率が急速に低下するのに対し，再現率の利得はわずかであり，実用的に用いるには，まだ不十分な精度である．

この大きな原因はコーパス量にあると思われる．提案手法で関係を抽出するためには，1 文内に少なくも 2 つの部位用語が共起する必要があるが，実験に用いたコーパスでは，400 余語の共起の組み合わせのごく一部しかカバーしきてれいない．よって，今後，より大きなコーパスを自動獲得する枠組みが求められよう．

また，限られたコーパスに対して，より深い処理を行い，有効利用することも考えている．例えば，提案手法の関係抽出パターンは，単純な語列のパターンであるが，これを部分木パターン [4] にすることなどは至近の課題である．

4 関連研究

一般的な関係抽出は MUC[2] や ACE[‖] のタスクにも採用され，様々な手法が提案され，試されてきた．

先行研究との差異は，関係タイプ間の関係（メタ関係）を利用している点である．例えば，ACE の Relation Detection & Characterization(RDC タスク) では，AT, NEAR, PART, ROLE, SOCial など 5 つの関係が取り扱われている．これらは，ほぼ独立した関係なので，個別に関係抽出を行うしかない．我々は人体部位用語にしぼって複数の関係タイプを扱うため，ある関係タイプで得られた情報を他の関係タイプに適応して，精度を高めることができる．我々の知る限り，本研究のように関係タイプ間の情報に着目した関係抽出の研究はない．

[‖] http://www.itl.nist.gov/iaui/894.01/tests/ace/

表 2: パターン “a における b” から得られる PART-OF 関係の例

	得られた PART-OF 関係とそのコンテキスト
	.. 脳の 脊髄 における 灰白質 と..
	.. 開放した 延髄 における 脳神経 の一般的な..
	..眼窩 における 頬骨 は蝶形骨大翼眼窩面..
√	..関節 における 足 の屈伸..
√	ある 関節 における 筋 の作用は、その..
√	.. 原因として 下肢 における 骨 の先天奇形や..

* √ は STEP2 にて棄却された関係を示す..

5 まとめ

本稿では複数の種類の関係を抽出する際に関係同士の関係（メタ関係）を用いる手法を提案した．また，小規模な実験によって，その枠組みがうまく働くことを検証した．今後は，より大きなコーパスを用いて実験を行い，実証的に手法の妥当性を検討する予定である．

参考文献

[1] Sergey Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, 1998.

[2] Ralph Grishman and Beth Sundheim. Message understanding conference- 6: A brief history. In *Proceedings of International Conference on Computational Linguistics (COLING1996)*, pp. 466–471, 1996.

[3] Hatano K. and Ohe K. Information retrieval system for japanese standard disease-code master using xml web service. In *American Medical Informatics Association (AMIA) Symposium*, p. 597, 2003.

[4] Ralph Grishman Kiyoshi Sudo, Satoshi Sekine. An improved extraction pattern representation model for automatic ie pattern acquisition. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2003)*, pp. 224–231, 2003.

[5] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of japanese morphological analyzer juman. pp. 22–28, 1994.

[6] SNOMED. *SNOMED Clinical Terms Guide*. College of American Pathologists, 2002.