電子カルテからの医療用語抽出

Extracting Clinical Terms from Discharge Summaries

石原健太　山本幸也　Davy Weissenbacher　佐々木裕
Kenta Ishihara　Yukiya Yamamoto　Davy Weissenbacher　Yutaka Sasaki

豊田工業大学
Toyota Technological Institute

1 はじめに

医療・医学分野において，膨大な量の文書情報が利用可能になるのに伴い，テキストマイニング技術が注目を集めている．しかしながら，テキストマイニングにとって分野文書は必須の情報であるにもかかわらず，患者のプライバシー保護に関する制約により，文書情報の入手が医療テキストマイニング研究のボトルネックとなってきた．

その意味で，2010 年に開催された i2b2 (Informatics for Integrating Biology and the Bedside) チャレンジは実際の医療の場面で作成された英語の電子カルテの情報にアクセスするための貴重な機会である．i2b2 チャレンジのオーガナイザは，349 件の訓練用電子カルテ (discharge summary) と 377 件のテスト用のカルテに専門家によるアノテーションを付与して，チャレンジの参加者に提供した．

文書の件数は比較的少ないが，先に述べたボトルネックのため，大量の医療文書の研究用利用は元来困難であり，いかに少ない文書から性能の高いシステムを構築するかが，医療分野におけるテキストマインニングの研究課題とも言える．本論文の共著者を含めてチャレンジの参加者全員が，米国 NIH が提供する被験者実験倫理に関する Web コースを受講し，修了証を得ることが義務付けられた．

2010 年の i2b2 チャレンジは医療概念抽出 (Concept Extraction)，表明分類 (Assertion Classification)，関係認識 (Relation Identification) の３種類のタスクからなり，本論文では，前者２つのタスクに関する研究結果について述べる．

2 i2b2 チャレンジの概要

i2b2 チャレンジでは毎年医療情報に関するタスクを出題しており，2010 年においては，電子カルテから重要情報をカテゴリ分類し抽出するタスクが与えられた．電子カルテは図１のような形で与えられている．

```
...
PRINCIPAL PROCEDURE :
5/5/05 right parietal occipital
craniotomy and debulking of tumor
using …
...
```

図１　電子カルテの例

以上の様な電子カルテから i2b2 オーガナイザが規定したアノテーションガイドラインに則り，医療に関する用語，表明，関係を抽出することが目標である．どのタスクもあらかじめトークン分け (tokenize) されたプレインテキストと訓練用のアノテーション情報が与えられた．コーパスのアノテーション数に関する情報は表１のようになる．

概念抽出タスク[1] は次の３種類の医療用語を抽出するタスクである．対象は，*Problem* (病気，怪我など)，*Treatment* (治療，阻害剤，病状を改善する物質)，および *Test* (医療上の問題を判定または除外するための手続き)である．例えば図１の

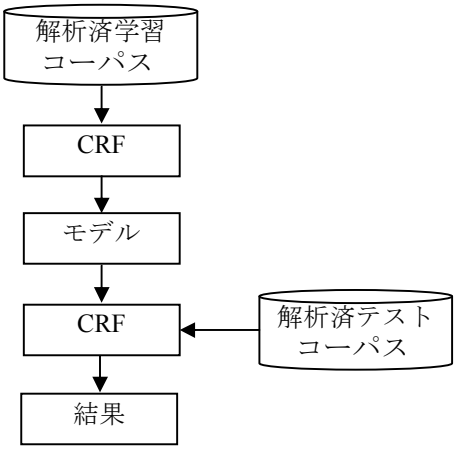

図 2　概念抽出のブロック図

例では debulking of tumor を *Treatment* として抽出する.

表明分類タスク[2] は, *Problem* に分類された用語がどのような文脈で表明されているかを分類する課題である. 分類対象は, 以下の 6 種類である.

(1) *Present*：患者が現在患っている
(2) *Absent*：患者が現在患っていない
(3) *Possible*：引き起こされる可能性のある
(4) *Conditional*：条件がそろった時に引き起こされる,アレルギーなど
(5) *Hypothetical*：仮説的にあげられる症状
(6) *Associated With Someone Else*：患者には直接無関係

表明分類タスクは, 概念抽出タスクの結果に対する処理を行うタスクであり, 前段の概念抽出に関する正解も入力として与えられる.

関係認識タスクは, 医療概念間の関係を認識するタスクである. 本論文では対象としていない.

3. 概念抽出タスク

3.1 概念抽出の概要

図 3 は概念抽出システムのブロック図である. 本研究と近い固有表現抽出においては(1)パターンマッチを用いる方法, (2)機械学習を用いる方法の 2 種類が大きく分けて存在する. 専門家がルールをすべて書き出せる場合は(1)を用いる方法は有効的だが, 今回は専門家が含まれていないので, 機械学習を用いる. 具体的には, CRF[3]を用いて医療用語の抽出を行う. CRF ツールとしては, CRF++[4]を利用している.

図 3 における解析済みコーパスは, コーパスに形態素解析を行った後, 単語の素性を付与したものを指す. 単語に付与した素性は. 「POS」「Chunk タグ」「Protein type」「最出文字種類」「接尾辞（3 文字）」「全文字種類」「tf-idf 値」「Wikipedia Category」の 8 種類ある. 先述の順で以下はその例になる.

	Training	Test	Total
#documents	349	377	726
# annotations	27,837	45,009	72,846
Test	7,369	12,899	20,268
Treatment	8,500	13,560	22,060
Problem	11,968	18,550	30,518
Present	8,052	13,025	21,077
Absent	2,535	3,609	6,144
Possible	535	883	1,418
Hypothetical	651	717	1,368
Conditional	103	171	274
Unassociated	92	145	237

表 1. コーパスのアノテーション数

```
movement movement NN I-NP O W ent W
*** NoCategory
```

図 3　単語の特徴例

形態素解析は GENIATagger を用いて行い, 「POS」「Poetry form」「Protein type」はその結果得られる. また「最出文字種類」「接尾辞（3 文字）」「全文字種類」「tf-idf 値」は形態素解析された単語から, そして, 「Wikipedia Category」は BGV 法[5]を用いて単語の尤もらしいカテゴリを抽出し特徴として付与した.

表 2　単語のみを素性とした場合（ベースライン）

種類＼値	再現率	適合率	F 値
Problem	0.6169	0.8603	0.7186
Treatment	0.5691	0.8846	0.6927
Test	0.6689	0.8846	0.6927
合計	0.6174	0.8828	0.7206

表 3 3.1 節の素性で学習を行った場合

種類＼値	再現率	適合率	F 値
Problem	0.6718	0.8580	0.7560
Treatment	0.6472	0.8951	0.7513
Test	0.7207	0.9224	0.8092
合計	0.6800	0.8873	0.7700

表 4　表 1，2 の差分

種類＼値	再現率	適合率	F 値
Problem	0.0549	-0.0023	0.0374
Treatment	0.0781	0.0105	0.0586
Test	0.0518	0.0378	0.1165
合計	0.0626	0.0045	0.0494

3.2 概念抽出の実験結果

表 2，3，4 に評価結果を示す．表 2 は単語のみを素性としたベースラインの結果である．表 3 は 3.1 節で述べた素性とコンセプトごとに学習を行う方法を採用した場合の結果である．表 4 に各項目の差分の再現率，適合率，F 値を示す．その結果 4.94%の精度向上が確認できた．

3.3 概念抽出に関する考察

概念ごとに結果を見たとき，それぞれの結果にばらつきがあることがわかった．つまり，概念に共通の素性を付与していたが，それぞれに適合した概念が必要になることがわかる．特に *Problem* は適合率が下がっており，すべて同じ特徴を付けるだけでは全体で良い結果でも，部分的に下がるので，個別に素性を検討する必要性がある．

4．表明分類タスク

4.1 表明分類の概要

表明の分類にも CRF を用いた．CRF を適用するために，電子カルテの各単語に特徴付けを行い，学習と分類を行った．素性として以下のものを使用した．

①分類する単語かどうか
　分類が必要な単語かを素性とした．
②概念的な単語の素性
　Genia Tagger [6]を使用し，品詞付と chunk を素性とした．
③表層的に決まる単語の素性
　使用した文字種，単語の型，原型，単語の前 4 文字，単語の後ろ 3 文字
④文脈からの素性
　以上の素性をその単語の前後 3 つすべてを文脈からの素性とした．

4.2 表明分類の実験結果

先に述べた素性を使い，CRF での実験の結果を図 4 と表 5 に示す．評価尺度として F 値を用いた．素性の組み合わせは以下の通りである。

system1　①　（ベースライン）
system2　①＋②
system3　①＋②＋③
system4　①＋②＋③＋④

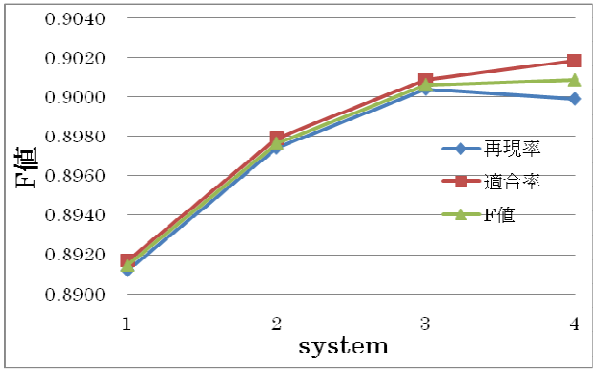

図 4　表明の実験結果のグラフ

表５　表明の分類結果

	再現率	適合率	F 値
system1	89.12%	89.17%	89.15%
system2	89.74%	89.79%	89.77%
system3	90.04%	90.09%	90.06%
system4	89.99%	90.18%	90.09%

表６　表明分類の結果詳細

	再現率	適合率	F 値
Present	98.16%	90.42%	94.13%
Absen	85.15%	92.98%	88.89%
Possible	33.30%	78.82%	46.82%
Hypothetical	68.20%	77.87%	72.71%
Conditional	16.96%	78.38%	27.88%
Unrelated	15.86%	85.19%	26.74%
all	89.99%	90.18%	90.09%

以上の結果より，system4 ではベースラインよりも F 値が 0.9 ポイントほど上昇していることがわかる．次に system4 の分類の結果を表 6 に示す．

５．おわりに

本論文では，医療文書からの用語・関係抽出をテーマとした 2010 i2b2 チャレンジの概念抽出，表明分類のタスクに関する研究結果について述べた．訓練データから CRF を用いてモデルを学習することにより，概念抽出では 77.00%，表明分類では 90.09% の F 値を達成することができた．

概念抽出に関する今後の方針として一つは，各特徴をコンセプトごとに見直し，最適な素性を発見する必要がある．例えば Wikipedia から抽出した Category は Wikipedia が定める Main Category を用いていたが，コンセプトごとに Sub Category を検討するなどの方法が考えられる．

また最適な素性を見つける以外にも，UMLS などの医療用語辞書を用いて意味的な素性を加えるなど，新たな素性の発見を行うなどの方法も考えることができる．

概念抽出，表明分類の双方において，素性を加え，性能を測定するという過程以外に，本研究において非常に時間がかかったのは素性の発見であった．今後，別のタスク以外に本手法を適用することもあるが，タスクによっては有用でない素性もあることは十分に考えられ，タスクが変わるごとに素性の発見を行なう場合，多くの時間を費やさなくてはならない．そこで今後の発展として，素性選定や上述の様なコンセプトごとに適合した素性の提案のための新しい手法の構築を検討していきたい．同様の手法を用いることで一定の水準まで性能を高めることが可能になり，広い分野で用いることが可能になると考えられる．

謝辞

本研究は豊田工業大学の平成２２年度特別研究費の支援を受けて実施した．

参考文献

[1] 2010 i2b2/VA Challenge Evaluation Concept Annotation Guidelines (http://www.i2b2.org/NLP/Relations/assets/Concept Annotation Guideline.pdf)

[2] 2010 i2b2/VA Challenge Evaluation Assertion Annotation Guidelines (http:/www.i2b2.org/NLP/Relations/assets/Assertion Annotation Guideline.pdf)

[3] J. Lafferty et al., Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ICML,2001

[4] CRF++, (http://crfpp.sourceforge.net)

[5]白川真澄ら『Wikipedia のカテゴリ構造解析とクラスタリングによる概念ベクトルの生成』JSAI 2009,2009

[6] Genia Tagger, (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/)