Self-Training を用いた電子カルテからの関係抽出

Relation Extraction from Discharge Summaries with Self-Training

大場弘樹　　佐々木裕
Hiroki Oba　　Yutaka Sasaki

豊田工業大学
Toyota Technological Institute

yutaka.sasaki@toyota-ti.ac.jp

1 はじめに

近年の技術進歩により，医療分野において診療情報の電子化が進み非常に多くのデータが蓄積されることとなった．電子カルテの普及による効果は種々あるが，従来の電子カルテの仕様による医師の負担は未だ解決されていない[1]．そこで，本研究では自然言語によって書かれたテキストからの情報抽出の自動化を目指し，医療用語の関係抽出を行う．

関係抽出の方法としては，機械学習を用いて，あらかじめ規定された関係カテゴリに対して多クラス分類を行う．

なお，学習用データおよび評価用データとして2010年に開催された i2b2 (Informatics for Integrating Biology and the Bedside)チャレンジ[2]に用いられたデータを使用した．このチャレンジは，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，本研究では関係認識のみを研究対象とする．関係のカテゴリは i2b2 2010で定義された８種類の関係を使用した．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．関係抽出のテストフェーズにおいて，前段階の医療概念抽出，表明分類の結果についは，タスクの仕様に従い正解データを使用した．

2 分類手法

本研究で抽出する関係は，医療の専門家により選定された症状，治療，検査の間の２項関係である．

- TrIP(治療が症状を良化させた)
- TrWP(治療が症状を悪化させた)
- TrCP(治療が新たな症状を発生させた)
- TrAP(症状に対し結果の示唆されていない治療)
- TrNAP(他の症状のために実行されなかった治療，又，同様な理由で中断された治療)
- PIP(症状が他の症状を示唆している)
- TeRP(検査が症状を明らかにした場合)
- TeCP(症状に対し結果が示唆されていない検査)

分類問題を扱うにあたって，様々な手法が考えられるが，本研究では SVM を用いる．SVM を用いた理由としては，分類問題に対して非常に高精度であり，特徴ベクトルの次元を増やしても精度が落ちない点が挙げられる．また，SVM は本来二値分類器であるが，関数距離が最大となるクラスを正例とする one-against-the-rest 法を用い，多クラス分類器に拡張する．

さらに関係抽出性能を向上させるため，半教師あり学習法として Self-Training を使用した．しかし，半教師あり学習に用いるデータは医療用語の抽出が為されていないため，過去に我々が作成したモデル[3]を

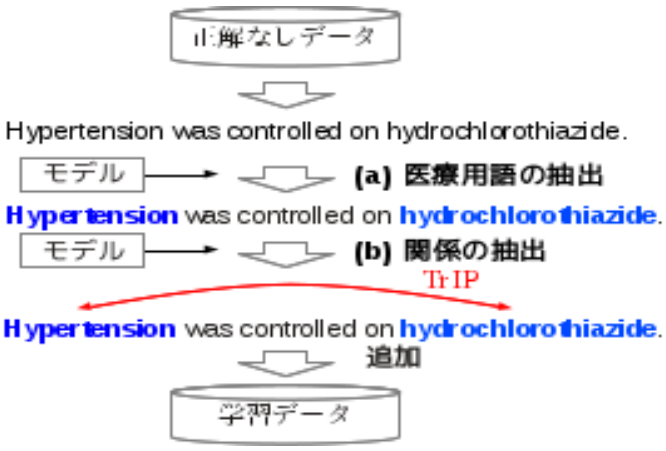


図１：Self-Training のプロセス

使用し医療用語の抽出を行っている．また，なるべく正確なデータのみを正解データに加えるようにするため，図１に示すような Self-Training のプロセスを行った．学習データに加える関係に関して，図１の(a)の用語抽出の段階で閾値を設定し，図１の(b)の段階でも閾値を設定している．図１の(a)で使用したモデルの精度を表１に示す．

表１：医療用語抽出モデルの精度

	Recall	Precision	F-score
All	0.7922	0.8400	0.8154

３　素性選択

以下に実験に用いた素性をまとめる．

(a) 同文中に出現する２つの医療用語(以下，解析対象)
(b) 解析対象の前後の単語
(c) 解析対象の前後の n-gram
(d) 解析対象の前後の品詞，句
(e) 単語間距離
(f) 症状の種類
(g) 解析対象と係り受け関係にある語
(h) 構文間距離

解析対象の前後の単語，品詞，句の数については，それぞれ最も効果的であった(単語数, 品詞数, 句数)=(2, 2, 4)を素性とした．同様に n-gram については，trigram まで効果が見られ，最も効果的であった(bigram 数, trigram 数)=(2, 1)を選択した．症状の種類については，文脈情報のタスクで用いられた正解データを使用した．単語間距離，構文間距離は一つの素性として与え，距離の逆数を素性の重みとした．品詞，句情報の獲得には，GENIA tagger を用い，係り受け解析には，Enju を用いた．

UMLS 等の専門辞書といった言語資源は一切用いることなく，訓練データから得られる素性のみを用いている．

４　実験と評価

使用したデータを表２にまとめる．

表2．使用したデータ

	学習(文)	テスト(文)
全体	5264	9070
TrIP	107	198
TrWP	56	143
TrCP	296	444
TrAP	1423	2487
TrNAP	106	191
PIP	1239	1986
TeRP	1734	3033
TeCP	303	588

アノテーションなしデータ：62,269文

評価には Recall，Precision を用いる．

$$\mathrm{Re}call = \frac{\text{システムが解答した内の正解数}}{\text{テストファイル全体の正解数}}$$

$$\mathrm{Pr}ecision = \frac{\text{システムが解答した内の正解数}}{\text{システムの解答数}}$$

また，一般に再現率と適合率はトレードオフの関係にあり，両方の評価の尺度として下に示すF値を用いる．

$$F\text{値} = \frac{2}{\frac{1}{\mathrm{Re}call} + \frac{1}{\mathrm{Pr}ecision}}$$

ベースラインを，解析対象の医療用語のみを素性として与えたものとし，その結果を表3に示す．3節で示した素性を全て用いて得られた結果を表4に，また，表4で得られた結果より，各素性を使用しない場合どれだけF値が落ちるかを表5に示し，各素性の効果について考察する．また，表6に半教師あり学習を用いた結果を示し，閾値を変動させた場合の結果を図2に示す．

表３：ベースライン

カテゴリ＼値	Recall	Precision	F値
TrIP	0.0202	0.8000	0.0394
TrWP	0.0000	0.0000	0.0000
TrCP	0.1104	0.8909	0.1964
TrAP	0.0523	0.5882	0.0960
TrNAP	0.0209	0.5714	0.0404
PIP	0.0081	0.6957	0.0159
TeRP	0.9314	0.5644	0.7029
TeCP	0.0085	0.1786	0.0162
ALL	0.3344	0.5676	0.4208

表４：３節に示す素性を全て用いた結果

カテゴリ＼値	Recall	Precision	F値
TrIP	0.2071	0.6212	0.3106
TrWP	0.0490	0.3684	0.0864
TrCP	0.4054	0.5788	0.4768
TrAP	0.6819	0.6968	0.6893
TrNAP	0.1309	0.4717	0.2049
PIP	0.5982	0.7110	0.6497
TeRP	0.8417	0.8196	0.8305
TeCP	0.3418	0.6955	0.4584
ALL	0.6495	0.7403	0.6919

表３，表４に示す結果より，全体でベースラインの F 値より約0.27向上したことがわかる．各カテゴリに対しても，全て向上している．

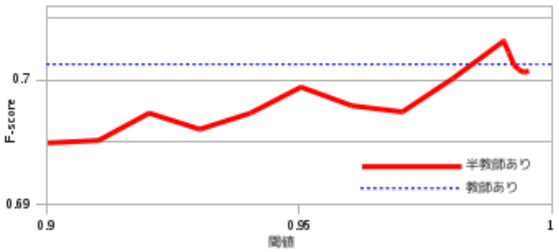
表5．各素性の効果

除いた素性	除く前との差
前後の単語	-0.0076
前後のngram	-0.0056
単語間距離	-0.0019
前後の品詞，句	-0.0018
症状の種類	-0.0091
係り受け関係にある語	-0.0191
構文間距離	-0.0021

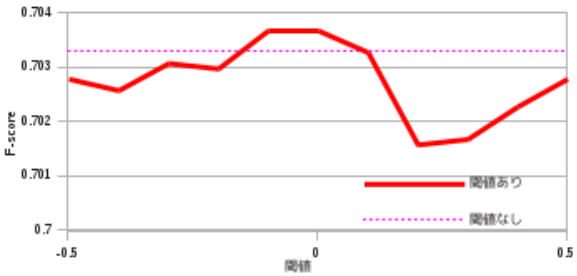
各素性の効果については，表５から係り受け関係にある語が最も効果的であることがわかる．また単語間距離と構文間距離においては片方のみを除いた場合、変化はあまり見られないが、両方を除くと F 値が0.08下落した．このことより，有効な素性であると考えられる．

表6：半教師あり学習の結果

カテゴリ＼値	Recall	Precision	F値
TrIP	0.2071	0.7455	0.3241
TrWP	0.0280	0.6667	0.0537
TrCP	0.4414	0.6144	0.5138
TrAP	0.7157	0.6809	0.6979
TrNAP	0.1414	0.5094	0.2213
PIP	0.6113	0.7116	0.6576
TeRP	0.8470	0.8428	0.8449
TeCP	0.3554	0.7061	0.4729
ALL	0.6659	0.7460	0.7037



(a)　図１(a)での閾値を変動させた場合



(b)　図１(b)での閾値を変動させた場合

図２：閾値の変化と半教師あり学習の性能

図２(a)より，閾値を小さく設定して教師あり学習データへ加えた場合，教師あり学習の結果を下回ることとなったが，閾値を上げていくと教師あり学習の結果を上回り，0.99で最も良い結果となった．図２(b)では，あまり大きな変化は見られず，これは(a)の段階で誤りを含む可能性があるデータを選別した結果，効果がはっきりと現れなかったのではないかと考えられる．実際に閾値0.99を通過したデータ数は6,817文しか通過していない．これについては、今後データを増やして再度確かめたい．

５ 関連研究

今回の実験では，全体でＦ値７割を達成した．この結果は他の i2b2チャレンジに参加チーム中の最上位システムと比較した場合，表7に示す通り，外部資源を使わないアプローチを採用した関係抽出システムとしては最も良い結果となった．

表7．i2b2-2010参加最上位チームとの比較

	Recall	Precision	Ｆ値
外部資源なし[4]	0.7496	0.6513	0.6970
外部資源あり[5]	0.7534	0.7204	0.7365
本研究	0.6659	0.7460	0.7037

６ まとめと今後の課題

i2b2 2010のデータを用いて関係抽出の研究を行った．Self-Training による半教師あり学習を導入することで関係学習の性能を向上させ，i2b2 2010に参加したシステムの中で外部資源を使わないシステムのカテゴリの中で最も性能の良いシステムよりも良いＦ値を達成することができた．

半教師あり学習のためのデータを増やすことは，教師あり学習用データを増やすのに比べて難しくないことが利点であるので，今後，他の関連する電子カルテデータを入手し，アノテーションなしデータを増やせないかを検討して行きたい．

参考文献

[1] 医療現場における電子カルテの影響 (http://ir.library.osaka-u.ac.jp/metadb/up/LIBKIYOK01/hs35-153.pdf)

[2] Relation Annotation Guidelines (https://www.i2b2.org/NLP/Relations/assets/Relation Annotation Guidelines.pdf)

[3] 電子カルテからの文脈情報抽出の性能向上法，2012年，豊田工大卒論，高橋竜二

[4] Jonnalagadda S, Gonzalez G., Can distributional statistics aid clinical concept extraction?, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, 2010.

[5] Roberts K, Rink B, Harabagiu S., Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/VA shared task, Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data, 2010.