臨床試験計画書の MeSH カテゴリーへの自動分類

Classifying Clinical Trial Protocols into MeSH Categories: Preliminary Report

佐々木 裕
Yutaka Sasaki

豊田工業大学
〒 468-8511 名古屋市天白区久方 2-12-1
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku, Nagoya 468-8511

1　はじめに

臨床試験（Clinical Trial）とは、新薬などの新しい治療法の開発にあたって、その治療法が人体に与える効果や影響を医学的に調査するための試験である。臨床試験は、第 I-IV 相（Phase I-IV）があり、順に商品化に近付いていく。臨床試験は国から治療の承認を得るための重要な過程であり、かつ実施には非常にコストと時間が掛かり、またリスクも存在するため、慎重に試験計画を設計する必要がある。

臨床試験を行なう際には、臨床試験を実施する組織が臨床試験計画書（Clinical Trial Protocol）を作成する。計画書においては、用いる薬や治療法、実施方法、実施期間、対象者の条件や人数等を、医学的及び統計学的見地から慎重に検討し、明確に記述する。

国連及び各国の医療保健機関が、臨床試験計画書の登録サイト（registry）を運営しており、どのような臨床試験が行われているかの情報が公開されている。アメリカにおいては、臨床試験に際しては、臨床試験計画書を政府の運営する登録サイト clinicaltrials.gov に登録することが義務付けられている。

臨床試験計画書の作成には、約 1 ヶ月程度の時間を要すると言われている。現在は、計画書の作成は、登録サイトに登録されている過去の臨床試験計画書や論文に記述されている臨床試験を人手で検索し、参考にしながら行なっている。本報告では、自然言語処理技術により臨床試験計画書の作成を支援することを大目標に、臨床試験計画書を、医学分野の概念シソーラスである MeSH（Medical Subject Headings）[3] カテゴリーに自動分類した結果について報告する。これにより、臨床試験計画書を MeSH カテゴリーに分類するためには、計画書に書かれているどの情報が有効であるかを明らかにする。

2　CTC コーパス

臨床試験計画書の分類実験のために、CTC コーパスを作成した。表 1 にコーパスの概要を示す。コーパスは下記の手順により作成した。

1. clinicaltrials.gov サイトより、登録されている計画書を HTML 形式ですべてダウンロー

表 1: CTC コーパスの概要

項目	値
文書数	82,525
コーパスサイズ	529MB
平均文書サイズ	6.4KB
カテゴリ総異なり数	11,443
トップレベルカテゴリ異なり数	63
平均カテゴリ付与数	25.9
平均トップレベルカテゴリ付与数	3.4

ド (2009.12.11 時点)[1]。

2. 同時に、MeSH カテゴリーリストを HTML 形式でダウンロード。

3. 計画書および MeSH カテゴリーを独自の XML 形式に変換。

2.1 計画書の構成

国連および各国の医療保健機関のレジストリに登録されている臨床試験計画書の項目は統一されていない。そのため、本実験では、clinicaltrials.gov サイトに登録されている下記の項目を実験対象とした。

- NCT ID: 計画書固有の ID
- Brief Title: 一般的なタイトル
- Official Title: 専門的なタイトル
- Brief Summary (Purpose): 臨床試験実施の概要
- Detailed Description : 臨床試験実施の詳しい説明
- Additional relevant MeSH terms: MeSH カテゴリー名リスト

MeSH カテゴリーは、各計画書に複数付与されているため、本分類問題は、多クラス多ラベル（multicalss multilabel）の分類問題となる。

2.2 MeSH カテゴリー

この節では、MeSH カテゴリーの概要について述べる。MeSH は、米国立衛生研究所（NIH: National Institutes of Health）の国立医学図書館（NLM: National Library of Medicine）が定める生物・医学用語のシソーラスである。

MeSH シソーラスのカテゴリー名（MeSH terms）には MeSH Tree Number という固有の ID が付与されている。この ID は、シソーラスの木構造における root ノードからの階層構造を表現した形式をとっている。

例えば、人間のインフルエンザは、下記の階層の下に分類されている。

- Virus Deseases [C02]
  - RNA Virus Infections [C02.782]
    - Orthomyxoviridae Infections [C02.782.620]
      - Influenza, Human [C02.782.620.365]

このように、ある MeSH カテゴリーの MeSH Tree Number を参照すれば、そのノードが継承している上位カテゴリーの ID を抽出できる。

clinicaltrials.gov サイトの MeSH カテゴリーの登録は、継承関係を持つノードを同じ計画書に割り当てることを許しているが、シソーラスの root から登録ノードに至るカテゴリがすべて登録さているとは限らない。本実験のために、すべての登録カテゴリーについて、root からその

[1]XML 形式でのダウンロードも可能であるが、XML 形式の場合、文の途中で改行されているため、パラグラフ単位で改行されている HTML 形式のデータをダウンロードした。

表 2: トップレベルカテゴリ一覧

	Tree number	文書数
1	D27	41,573
2	C04	21,314
3	C23	20,453
4	D02	14,192
5	D03	11,273
6	C20	11,161
7	C14	10,461
8	C10	10,374
9	C13	9,903
10	C12	9,195
...	...	...
59	D26	8
60	E07	2
61	G10	2
62	N02	1
63	E02	1

ノードに至るまでのカテゴリーの Tree Number を生成し、各計画書に割り当てた。

本実験では、実験期間の制約のため 63 種類のトップレベルのカテゴリーへの分類実験を行なった。トップレベルのカテゴリーの一覧を表 2 に示す。

3 実験と評価結果

3.1 実験手法

１０分割交差検定により評価を行なった。学習アルゴリズムには SVM^{perf} [1, 2] を用いた[2]。多クラス多ラベル問題を解く必要があるため、各カテゴリに対して訓練データから one-against-the-rest 法 [5] による 2 値分類 SVM モデル[3]を学習し、テストデータにより評価した。評価尺度には micro-average F1 値を用いた。

[2]パラメータ c の値は 70 を用いた。

[3]そのカテゴリに計画書が属するか属さないかを判定するモデル。

3.2 素性

“Brief Title”, および “Official Title”, “Brief Summary”, “Detailed Description” の 4 つの記述項目に現れる単語を素性として用いた。単語はすべて小文字による表現に変換した。

本実験では、以下の 6 種類の素性の構成法を比較する。

a) ベースライン：すべての単語を記述項目を区別せず素性とする。

b) Brief Title の出現単語のみを素性とする。

c) Official Title の出現単語のみを素性とする。

d) Brief Summary の出現単語のみを素性とする。

e) Detailed Description の出現単語のみを素性とする。

f) 単語が出現した記述項目毎に、単語を区別しながら、すべての単語を素性とする。

素性の値は、単語の出現の有無を 0/1 で表現した値を用いた。

3.3 結果

実験結果を表 3 に示す。全体的な傾向として、カテゴリ数が増えるにしたがって、性能は低下していく。これは、出現回数の少ないカテゴリーについては訓練データがスパースになるため、性能が下がると考えられる。また、記述項目別では、Brief Title のみを素性とした場合の F 値が最も高かった。全体的に良い性能が出ているのは、対象として 4 項目のすべての単語を出現項目を区別した素性として扱った場合である。これは、一見当然の結果であるが、臨床試験計画書についてこの点を初めて明らかにした点には意義があると考える。また、臨床試験計画書の

表 3: 実験結果

	頻度トップ n カテゴリー					
素性タイプ	10	20	30	40	50	63
a) ベースライン	81.30	81.06	80.26	79.46	79.25	79.21
b) Brief Title	80.89	80.52	79.48	78.78	78.55	78.51
c) Official Title	79.71	79.43	78.48	77.82	77.62	77.59
d) Brief Summary	79.91	79.53	78.59	77.84	77.61	77.57
e) Detailed Description	64.91	62.24	61.05	60.25	60.09	60.04
f) 記述項目別	83.77	83.62	82.97	82.39	82.23	82.18

（micro-average F1 スコア）

MeSH カテゴリーへの自動分類の研究自体が過去に行われておらず、臨床試験計画書の素性の有効性についての特性を明らかにすること自体がチャレンジングなテーマである。

4 まとめ

本報告では、臨床試験計画書の MeSH カテゴリーへの自動分類に関する基本的な実験とその結果について述べた。臨床試験計画書の分類実験のために、新たに CTC コーパスを構築し、コーパスに出現する最上位レベルのカテゴリー 63 種類について分類実験を行った。実験には 10 分割交差検定を採用した。SVM^{perf} により、各カテゴリに対して訓練データから one-again-the-rest 法による 2 値分類 SVM モデルを学習し、テストデータにより評価した。その結果、出現数上位のカテゴリーついては、記述項目を考慮しながら単語を素性とすることで、F 値 82 を越える性能が得られることが明らかとなった。今後、最上位階層以外の階層のカテゴリーについても分類性能を評価していきたい。そのためには、階層関係を利用しながら、大量のカテゴリーを多クラス多ラベル問題として効率良く解くことができるかが鍵となると考えている。

参考文献

[1] Thorsten Joachims, A Support Vector Method for Multivariate Performance Measures, *Proc. of the International Conference on Machine Learning (ICML-05)*, pp. 377–384, 2005.

[2] Thorsten Joachims, Training Linear SVMs in Linear Time, *Proc of the ACM Conference on Knowledge Discovery and Data Mining (KDD-06)*, pp. 217 - 226, 2006.

[3] Medical Subject Headings, http://www.nlm.nih.gov/mesh/

[4] MeSH Tree Number, http://www.nlm.nih.gov/cgi/mesh/2010/MB_cgi

[5] Jason Weston and Chris Watkins, Support Vector Machines for Multiclass Pattern Recognition, *Proc. of the Seventh European Symposium on Artificial Neural Networks*, Brussels, pp. 219-224, 1999.