臨床試験論文アブストラクトに対する情報抽出および文分類の適用

原　一夫　　松本裕治

奈良先端科学技術大学院大学

{kazuo-h, matsu}@is.naist.jp

1 はじめに

近年、医療現場では Evidence-Based Medicine（EBM）という概念が急速に普及している。EBM とは、Sackett ら[1]によれば、“ the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients ”であり、医療を行う者は最新で最良のエビデンスに基づいた治療方法を常に把握している必要があるということを意味する。エビデンスに基づく医療のための情報源としては、アメリカ国立医学図書館（NLM: National Library of Medicine）の生物工学情報センター（NCBI）が管理する MEDLINE があり、多数の医薬生物学系論文アブストラクトが利用可能である。しかし、膨大な数の論文アブストラクトを格納するデータベースから必要とする情報を利用者が見つけ出すのは必ずしも容易なことではない。そこで、我々は MEDLINE の臨床試験論文アブストラクトに焦点をあて、エビデンスに基づく医療を実践するために必要となるであろう情報の自動抽出を試みた。

一方、近年の自然言語処理の研究において、情報抽出（IE）は多くの研究者が取り組んでいる課題である。たとえば 1987 年から 1998 年までアメリカで開催された MUC（Message Understanding Conference）では、主催者が提示する特定の情報抽出タスクに対し、参加した研究者の多くは辞書と抽出パターンを手作業により工夫して作成することで抽出精度を競った。最近では、手作業によらず自動的に辞書および抽出パターンを生成することに関心が移ってきている。

本稿では、今回の情報抽出の対象である「治療方法」、「評価項目」、「対象患者」について 2 節で簡単に述べてから、それらに対応する名詞句を臨床試験論文アブストラクトから抽出する情報抽出実験の結果について 3 節で報告する。この実験結果から、我々はこの問題を文分類タスクと情報抽出タスクに分割することとした。すなわち、抽出したい情報を含んでいるであろう文をまず文分類タスクによって選択し、それらに対してのみ情報抽出タスクを適用することで、抽出したい名詞句を獲得する精度を高めることができるかもしれない。このことを踏まえ、我々は最新の文分類の手法を臨床試験論文アブストラクトに対して適用し、文分類実験を行った。この実験結果については 4 節で議論する。なお、本稿における実験の概要を図 1 に示す。

2 情報抽出の対象

臨床試験の目的は、ある医療、たとえば新薬候補化合物を用いた新しい治療方法が、それを実際の患者に適用するとき、どの程度の臨床上の有効性と安全性を示すかを調べることである。すなわち、医療の正当性を保証するエビデンスを客観的に明らかにすることである。臨床試験の結果を意味あるものにするには、試験を開始する前に試験の計画（治験デザイン）を十分客観的に設定しておく必要があるが[2]、今回情報抽出の対象とした「治療方法」、「評価項目」、「対象患者」は、臨床試験の手順書（プロトコル）や総括報告書に必ず記載されることであり[3]、臨床試験の治験デザインを要約する情報とも言える。

以下、具体例をあげると、「治療方法」は新薬候補化合物、対照薬、偽薬を用いた治療、もしくは薬剤を用いない外科的治療などである。「評価項目」は、血圧値、コレステロール値などの臨床検査値から、QOL（Quality of Life）改善度や生存率まで、臨床試験によって幅広く設定される。「対象患者」としては、成人男性なら誰でもよいという試験もあれば、開発中の薬剤の危険性を考慮して、細かな組入れ基準と除外基準を設ける試験もある。

3 情報抽出実験

3.1 方法

臨床試験論文アブストラクトのタイトルと本文に対し、TnT [4]により品詞タグ付けし、YamCha [5]を用いて名詞句のチャンキングを行う。そして切り出した名詞句に対し、我々が作成した表 1 に示すタグを用いてタグ付けする。このタグ付けは、機械学習の手法により自動化できるが、精度がいまのところ十分でないため、今回は人手に依る。最後に、表 3 に示す、手作業により作成した抽出パターンを用いて、「治療方法」、「評価項目」、「対象患者」に対する

情報抽出を行う。なお、アブストラクトのタイトルには重要な情報が含まれやすいため、抽出パターンはタイトル、本文に対し別々に作成している。

実験に用いたデータは、2004 年 10 月に PubMed（http://www.ncbi.nlm.nih.gov/entrez/query.fcgi）において、検索クエリーを “hepatitis [MeSH Terms] AND hasabstract AND Randomized Controlled Trial [ptyp]”[1] としてダウンロードした、MEDLINE への登録日付順に新しい 50 アブストラクトである。今回は、実験を簡単にするため、アブストラクトを肝炎に関係する臨床試験に限定する。

表 1: 名詞句に対するタグ

タグ	タグの適用範囲
DISEASE	disease, complication, symptom, virus
DRUG	drug, chemical compound, nucleoside, placebo
STUDY	clinical trial, statistical analysis
THERAPY	therapy, treatment, regimen, surgery
PATIENT	participants in the clinical trial
TARGET	endpoints, clinical laboratory evaluation
SCHEDULE	time schedule of the clinical trial
VALUE	value or score of TARGET
NUMBER	numeral expression

3.2 結果と考察

臨床試験論文アブストラクトのタイトルのみからの抽出結果、および、アブストラクト全体（タイトルと本文）からの抽出結果、それぞれについての precision、recall の値を表 2 に示す。ここで、重要な情報を密に含みやすいタイトルのみからの抽出結果は、今回の情報抽出実験のベースラインとも言える。なお、抽出対象となる名詞句（正答）は、それぞれのアブストラクトに対して複数存在することがあるが（たとえば、薬剤 A、薬剤 B という 2 つの薬剤を比較する臨床試験においては、「治療方法」に対する正答は 2 つ存在する）、そのすべてが抽出できた場合を recall (a)のカウントとし、その少なくとも一部が抽出できた場合を recall (b)のカウントとしている。

表 2 に示す実験結果は実用上必ずしも十分な結果とは言えないが、その主な原因は 2 つ考えられる。第一の原因は、手作業による抽出パターン作成の限界である。もちろん臨床試験の専門家はより良い抽出パターンを作成できるかもしれないが、こうした発見的方法では抽出方法の包括的な正当性は保証されない。そして第二の原因は、今回の情報抽出タスクは名詞句（あるいは固有表現）をその役割に応じて選択的に抽出するタスクであるということである。たとえば、「治療方法」を抽出するとき、アブストラクトに現れる薬剤名をすべて抽出して良いわけではない。著者が臨床試験の背景知識として記述しているにすぎない薬剤名は、当該臨床試験とは直接の関係を持たず、ゆえに今回の抽出対象とはならない。次節では、抽出対象となる名詞句（正答）を含む文は共通の性質（役割）を持っていると仮定し、その性質を持つ文をアブストラクトから選択、分類する実験の結果について報告する。

表 2: 情報抽出実験の結果

		治療方法	評価項目	対象患者
タイトルから	precision	81.5%	73.1%	93.8%
	recall (a)	40.0%	24.0%	48.0%
	recall (b)	92.0%	32.0%	60.0%
タイトルと本文（アブストラクト全体）から	precision	84.8%	77.0%	76.2%
	recall (a)	64.0%	52.0%	82.0%
	recall (b)	92.0%	64.0%	86.0%

4 文分類実験

4.1 方法

臨床試験論文アブストラクトの本文に対し、3 節と同様にして名詞句のタグ付けまでを行い、つづいて最新の文分類器の一つである BACT [6]を適用する。BACT は機械学習の手法を用い、トレーニング・データから最適な文分類方法を学習し、その方法をテスト・データに対して用いることで文分類を行うことができる。詳述すると、BACT が入力として受けつけるのは木構造（順序木）データであるが、トレーニング・データの木に対して BACT はそのすべての部分木を探索的に調べ、分類結果に強い影響を与える部分木を見つけ出す。そして BACT により自動抽出されたこれらの部分木は、テスト・データの木を分類するためのパターンを構成する。

今回の文分類実験では、単語の系列としての文に対し、3 通りの単語間関係、すなわち、係り受け、N グラム、bag-of-words（BOW）を仮定し、それぞれに対応する木構造に文を変換して BACT に適用する。ここで、係り受け関係に対応する依存構造木は、Charniak によるパーサー（句構造解析器）[7]の出力を Collins の主辞規則（head rule）[8]を用いて依存構造木に変換して作成する。

[1] この検索クエリーは、「ランダム化比較試験（Randomized Controlled Trial；信頼性の高い臨床試験の実施方法で、現在もっともよく採用されている）であり、かつ肝炎（hepatitis）に関するもので、かつアブストラクトのタイトルのみでなく本文が MEDLINE に登録されている（hasabstract）」という意味である。

入力文 ： 臨床試験論文アブストラクトからの例文

“We conducted a multi-center, randomized trial comparing peginterferon plus ribavirin with interferon plus ribavirin for the treatment of chronic hepatitis C in persons co-infected with HIV.”

名詞句のタグ付け

“[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] co-infected with [DISEASE].”

出力１ ： 情報抽出実験

治療方法: “peginterferon plus ribavirin”、 治療方法: “interferon plus ribavirin”、

評価項目: （この例文中には該当するものなし）、 対象患者: “persons co-infected with HIV”

出力２ ： 文分類実験（抽出対象を含んでいるかどうかによる文の分類）

治療方法: +1 (Yes) 、 評価項目: -1 (No) 、 対象患者: +1 (Yes)、 治験デザイン: +1 (Yes)

図 1: 本稿における実験の概要

表 3: 手作業により作成した情報抽出のためのパターン（情報抽出実験）

治療方法	手作業により作成した情報抽出のためのパターン
タイトルから	“DRUG” または “THERAPY” とタグ付けされた名詞句を抽出する。
本文から	以下の 2 通りの正規表現のいずれかにマッチする “DRUG” または “THERAPY” を抽出する； （1）“* DRUG (.* *as non-NP*) DRUG *” （2）“* (compare* | between) * (DRUG | THERAPY) * (versus | with | and) * (DRUG | THERAPY) *”
評価項目	手作業により作成した情報抽出のためのパターン
タイトルから	“TARGET”とタグ付けされた名詞句を抽出する。
本文から	以下の 2 通りの正規表現のいずれかにマッチする “TARGET” を抽出する； （1）“* (We | we) .* TARGET *” （2）“* ((This | this) * *as STUDY*) .* TARGET *”
対象患者	手作業により作成した情報抽出のためのパターン
タイトルから	“PATIENT” とタグ付けされた名詞句、もしくは以下の 2 通りの正規表現のいずれかにマッチする名詞句を抽出する； （1）“PATIENT * with DISEASE” （2）“PATIENT * with TARGET”
本文から	以下の 2 通りの正規表現のいずれかにマッチする名詞句を抽出する； （1）“PATIENT * with DISEASE” （2）“PATIENT * with TARGET”

表 4: BACT によって自動生成された“DRUG”を含むパターンとその重み（文分類実験）

BACT によって自動生成された文分類のためのパターンのうち “DRUG”を含むもの	治験デザイン	治療方法	評価項目	対象患者
文中で、“PATIENT” および “DRUG” が “received” に係る。	–	+ 0.048	–	–
文中に、“DRUG” とタグ付けされた名詞句が存在する。	+ 0.009	+ 0.046	–	–
文中で、“DRUG” が “of” に係り、“of” が “TARGET” に係る。	+ 0.009	–	+ 0.035	–
文中で、“DRUG” が “DRUG” に係る。	+ 0.008	+ 0.013	–	–
文中で、“DRUG” が “received” に係る。	–	+ 0.010	+ 0.023	–
文中で、“DRUG” が “of” に係る。	+ 0.011	+ 0.006	+ 0.012	–
文中で、“DRUG” が “with” に係る。	–	- 0.004	–	- 0.026
文中で、“DRUG” が “to” に係る。	- 0.004	- 0.013	–	- 0.012
文中で、“DRUG” が “in” に係る。	- 0.011	- 0.019	–	–

実験に用いたデータは 3 節の情報抽出実験と同じものであり、50 アブストラクトの本文（タイトルを除く）を構成する 562 文である。文の分類は、「治療方法」、「評価項目」、「対象患者」に対応する名詞句を含む（+1；yes）あるいは含まない（-1；no）により行う。さらに、「治療方法」、「評価項目」、「対象患者」のいずれかに対応する名詞句を含む文は「治験デザイン」を記述する役割を持つ文であると解釈し、この役割を持つ持たないによる文分類も行う。

4.2 結果と考察

文分類実験の結果を表 5 に示す。数値は 5 分割クロス・バリデーションによる F 値（precision と recall の調和平均）の平均値である。仮定した 3 通りの単語間関係による文分類結果に統計的な有意差（P>0.05）があるかどうかについてマクネマー検定（対応がある場合の比率の差の検定）を行い、有意な結果には表中の数値に下線を引いた（有意差は、いずれも BOW との比較で認められた）。

表 5: 文分類実験の結果

		治験デザイン	治療方法	評価項目	対象患者
文の総数		562	562	562	562
“yes” 文の数		160	90	76	55
BOW	F 値	77.4%	74.0%	74.3%	67.2%
N グラム	F 値	78.7%	75.5%	78.8%	<u>80.4%</u>
係り受け	F 値	<u>82.9%</u>	<u>81.6%</u>	77.8%	72.7%

仮に Charniak によるパーサーが十分正確に動作するならば、文を木に変換する際に多くの語彙知識を使っているため、係り受けは N グラムや BOW よりも良い結果を得ると考えるのが自然かもしれない。しかし、「対象患者」では N グラムが係り受けと比較して特に良い結果を得ている。この理由と思われるのは、Charniak パーサーによる前置詞句の係り先決定の誤り（PP attachment の問題）である。“ PATIENT with DISEASE ” という固定した表現を持つ文は、全 562 文のうち 71 文が存在するが、このうち「対象患者」を含む文は 29 文もあり、全正答数である 55 文の半分以上を占める。一方、“ with DISEASE ” の係り先をパーサーが正しく決定できていない文は 71 文のなかに 15 文（正答は 7 文）も存在する。よって、“ PATIENT with DISEASE ”という構造を壊すことのない N グラムが係り受けを凌ぐ結果になっていると考えられる。

次に、係り受けの依存構造木を入力とした場合の、BACT によって自動生成されたパターンについて考察する。表 4 に “ DRUG ” を含むパターンと、その文分類への寄与の重みを示す。数値の符号は、プラスは抽出対象を含むとする方向、マイナスは含まないとする方向への寄与を表し、絶対値は寄与の大きさである。ハイフンは BACT がそのパターンを抽出しなかったという意味である。この表によれば、“ DRUG ” を含むパターンは、文分類の結果を「評価項目」については含むとする方向へ、「対象患者」は含まないとする方向へと導く一方で、「治療方法」と「治験デザイン」ではどちらの方向へも導くパターンが存在する。このことは、3.2 節で述べたように、アブストラクトに現れる薬剤名は必ずしも「治療方法」の抽出対象とはならないことを反映している。

5 おわりに

本稿では、臨床試験論文アブストラクトから「治療方法」、「評価項目」、「対象患者」を抽出する情報抽出実験、および、それらを含む含まないによる文分類実験について報告した。手作業で作成した抽出パターンによる情報抽出実験では十分な結果は得られなかったが、その補助的タスクとしての文分類実験では、BACT が自動生成するパターンを利用すれば情報抽出の対象（正答）を含む文の絞り込みがある程度はできることを示した。ただし、BACT の入力として係り受けの依存構造木を用いる場合、PP attachment の問題などによるパーサーの誤りは解析精度向上の障害になる可能性を孕んでいる。本タスクに適した戦略の探究は今後も必要である。

参考文献

[1] Sackett et al.: Evidence based medicine: what it is and what it isn't. BMJ 312 (7023), 13 January, pp. 71-72 (1996).

[2] ICH E9 あるいは厚生省（当時）通知ガイドライン: 「臨床試験のための統計的原則」について(1998).

[3] ICH E3 あるいは厚生省（当時）通知ガイドライン: 治験の総括報告書の構成と内容に関するガイドラインについて(1996).

[4] Brants, T.: TnT - a statistical part-of-speech tagger, In Proceedings of ANLP, pp. 224-231 (2000).

[5] Kudo, T. and Matsumoto, Y.: Chunking with Support Vector Machines, In Proceedings of NAACL, pp. 192-199 (2001).

[6] 工藤拓, 松本裕治: 半構造化テキストの分類のためのブースティングアルゴリズム, 情報処理学会論文誌, Vol.45, No.9, pp. 2146-2156 (2004).

[7] Charniak, E.: A Maximum-Entropy-Inspired Parser, In Proceedings of NAACL, pp. 132-139 (2000).

[8] Collins, M.: Head-Driven Statistical Models for Natural Language Processing, PhD dissertation, University of Pennsylvania (1999).