4H1-OS-27a-3

文献データに基づく症例検索システムの構築

Development of a literature-based case discovery and case matching system

藤原 豊史[*1,2] Toyofumi Fujiwara　山本 泰智[*1] Yasunori Yamamoto　金 進東[*1] Jin-Dong Kim　高木 利久[*2] Toshihisa Takagi

[*1] 情報・システム研究機構 ライフサイエンス統合データベースセンター
Database Center for Life Science, Research Organization of Information and Systems

[*2] 東京大学大学院 新領域創成科学研究科 メディカル情報生命専攻
Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo

There are ~7,000 rare genetic diseases, and it is estimated that ~4% of newborns suffer from them. Recently, NGS-based diagnostic tests have been used to improve the diagnostic rate and early diagnosis for rare disease patients. In order to use the NGS-based diagnostic tests more effectively, physicians and researchers have shared and utilized case records. In particular, over 1 million case reports in PubMed play an important role in the investigation of rare diseases, and can be retrieved by using PubMed. However, there are many synonyms and many terms that are semantically related to each other in titles and abstracts, we need to correspond to them to efficiently search for case reports. To address the need for sharing case reports and searching for them, we developed a case matching system using multiple ontologies and evaluated it with ClinVar records.

1. はじめに

1.1 希少疾患患者の遺伝学的検査

希少疾患はおよそ 6,000-7,000 存在し[Orphanet 2017], その多くは遺伝性疾患であり, 新生児の約 4%は希少疾患に罹患している[Mutarelli 2014]. 現在, これら希少疾患の診断率向上および早期の診断が課題になっている. 例えば, 希少疾患患者の半数は生涯診断がつかないと報告されている[Sawyer 2016]. また, マルフォン症候群など希少疾患の中ではよく知られている 8 つの疾患の患者について診断状況を調査した結果, 約 25%は診断がつくまでに 5 年から 30 年を要し, 約 40%は最初の診断が間違っていた[EURORDIS 2007]. このような状況の中, 近年では, NIH の未診断疾患プログラム[UDP]や国立研究開発法人日本医療研究開発機構の未診断疾患イニシアチブ[IRUD]において, 希少疾患患者に Exome 解析などの遺伝学的検査が行われ, 診断率が高まっている[Sawyer 2016]. また Laurel らは, 遺伝性疾患の罹患が疑われる 4 ヶ月未満の新生児を対象に遺伝学的検査を実施し, その結果, 一部の患者に対して診断がつき, 早期に診断がついたことで極めて効果的な治療を選択できたと報告している[Laurel 2015].

しかし, 遺伝学的検査の適用には, さまざまな課題が存在する. 例えば, 遺伝学的検査には遺伝子パネル検査, 全エクソーム検査, 全ゲノム検査などが存在するが, 診断率を高めるためには患者の症状や病態に適した検査を選択する必要がある[Leslie 2014]. また, 前述の Laurel らは, 検査結果の解釈に要した日数の中央値は 23 であった(最短日数:5, 最長日数:912)と報告しているが, 診断を早めるためにはこの日数を短縮する必要がある.

これらの課題を解決するために既存症例が活用されている. 適切な遺伝学的検査の選択には, 対象とする患者の症状や病態が類似する既存症例が見つかれば, それら症例で適用された遺伝学的検査が参考になる. また, 検査結果の解釈にも, 患者の変異や症状・病態が類似する既存症例の診断結果が参考になる. しかし, 希少疾患はもともと症例数が少ないために, 効率的に既存症例を探すことが難しく, 既存症例の共有が国際的な課題になっている[Philippakis 2015].

1.2 既存症例の共有問題

現在, 医者や研究者が手動で症例を登録する「個別登録データに基づく症例検索システム」が数多く開発され(図 1), 表現型の異常に関する用語を集めた Human Phenotype Ontology (HPO) [Sebastian 2017]を用いてシステム間の情報共有も進められているが, 共有されてない情報も多く, それらを一元的に検索することは難しい.

一方で, 症例共有の手段として古くから利用されている症例報告は 100 万件以上存在し, PubMed を利用して一元的に検索することができる. また症例報告数は年々増加する傾向にあり, 特に希少疾患に関する症例報告は, 疾患の原因や兆候, 新たな治療の選択肢などの発見に役立ち, その重要性が報告されている[Carey 2010].

しかし, PubMed が検索対象とするタイトル・アブストラクトには類義語が多く含まれるため, 効率的な検索が難しい. また意味的に関係が深い用語が数多く存在し, 患者の症状・病態が類似する症例報告を効率よく検索するためには, 用語の意味的な関係性を考慮した検索が必要となる.

1.3 文献データに基づく症例報告検索システムの構築

PubMed に含まれる症例報告を対象とし, 類義語に対応し, かつ意味的検索が可能な「文献データに基づく症例検索システム」を構築した(図 1). 類似症例報告検索手法として, 複数オントロジーを用いる手法を採用した. また, HPO のみを用いる手法, 文章に含まれる単語をその位置を考慮せずにベクトル化する Bag-of-Words を用いた手法, 用語の周辺の文脈情報や文章全体の情報を考慮してベクトル化する Distributed Memory Model of Paragraph Vectors(以降, PV-DM と呼ぶ)を用いた手法と, 類似症例報告の検索精度を比較した. 検索精度を比較する

連絡先:藤原 豊史, 情報・システム研究機構 ライフサイエンス統合データベースセンター, 東京大学大学院　新領域創成科学研究科 メディカル情報生命専攻,
fujiwara @dbcls.rois.ac.jp

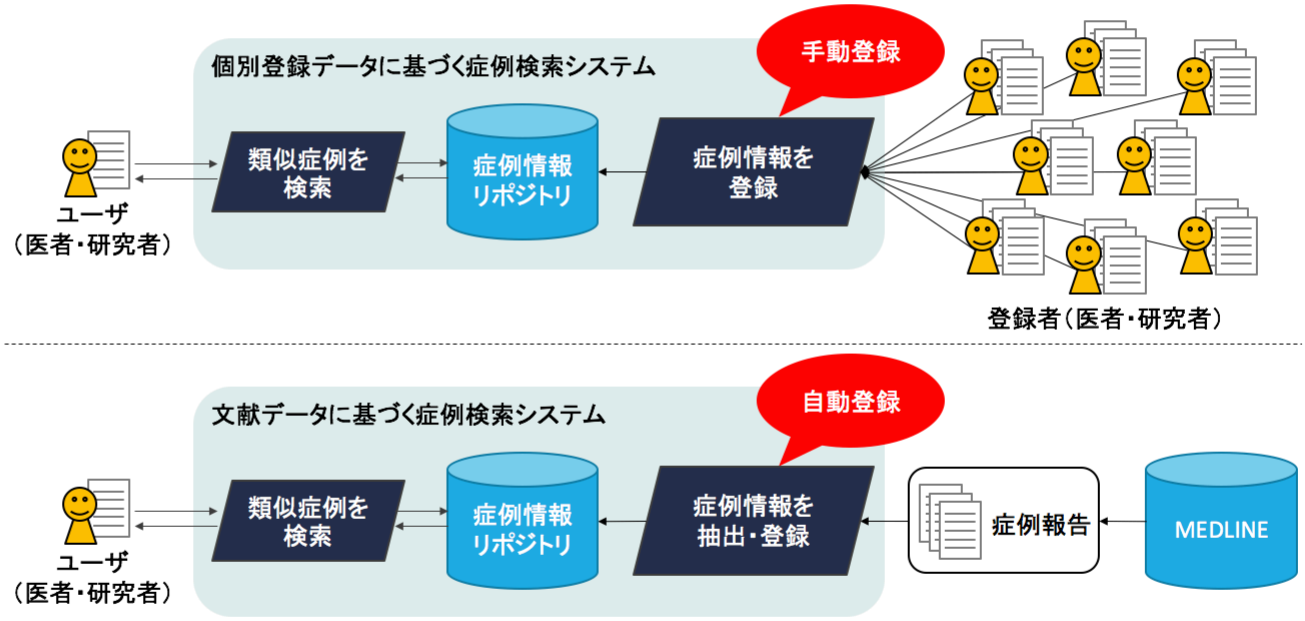

図 1.「個別登録データに基づく症例検索システム」と「文献データに基づく症例検索システム」の概要

ために，ClinVarp[ClinVar 2017]から遺伝子変異と疾患が共通する症例報告セットを取得し，評価用データセットを作成した．

2 章では類似症例報告検索手法について述べ，3 章では ClinVar を用いた評価用データセットの作成と評価結果を報告する．そして 4 章で今回の調査で明らかになった今後の課題について述べる．

2. 類似症例報告検索手法の構築

2.1 では検索対象となる症例報告の取得方法について述べ，2.2 では類似症例報告を検索する際に利用するアノテーションの付与方法ついて述べる．2.3 では，複数の類似症例報告検索手法について述べる．

2.1 PubMed を用いた症例報告の取得

PubMed[PubMed 2017]を用いて症例報告を取得するには，症例報告に割り当てられた”Case Reports”タグを利用する．このタグは，PubMed の索引作成者がジャーナル毎の形式などを判断材料として手動で割り当てている．また，文献タイトルに”case report”または”case reports”を含む文献も症例報告として取得する．以下の PubMed クエリを用いて，症例報告リストを取得した．

- "case reports"[Publication Type] OR "case reports"[ti] OR "case report"[ti]

検索の結果，約 180 万件の症例報告のリストを取得し，それらのタイトルおよびアブストラクトを PubMed から取得した．

2.2 オントロジーを用いたアノテーション付与

(1) 情報抽出に用いたオントロジー

オントロジーは，あるドメインの知識を概念化したものであり，is-a や part-of などで概念間の関係が定義され，機械可読形式で表現されている．オントロジーの各概念には，それらを表現するためのラベルが付けられており，あるテキストに含まれる情報（オントロジーの概念）を抽出する際に利用することができる．

我々は，多くの症例検索システムで利用されている HPO の他に，NCBO BioPortal[BioPortal 2017]に登録されている 540 件（2017 年 3 月時点）の生物医学オントロジーの中から選択した複数のオントロジーを利用した(表 1).

オントロジー名	概要	用語数
Human Phenotype Ontology (HPO)	ヒトの表現型の異常に関する用語集	11,877
Orphanet Rare Disease Ontology (ORDO)	ヒトの希少疾患名に関する用語集	8,910
Human Disease Ontology (DOID)	ヒトの疾患名に関する用語集	9,342
SNOMED CT	世界で最も包括的で正確な医療用語集	320,338
Foundational Model of Anatomy (FMA)	ヒトの解剖学に関する用語集	104,258
Mammalian Phenotype Ontology (MP)	哺乳類の表現型に関する用語集	11,856

表 1. 情報抽出に用いたオントロジー

(2) オントロジーを用いた情報抽出システム

オントロジーを用いた情報抽出システムは複数存在するが，ライフサイエンス分野の中では NCBO Annotator[BioPortal 2017]，MetaMap[Taboada 2014]，ConceptMapper[Groza 2015]がよく利用されており，また我々(ライフサイエンス統合データベースセンター)は PubDictionaries[PubDictionaries 2017]を開発している．この中から症例報告検索システムに最適なシステムを選択するために，処理速度と情報抽出精度を比較した．情報抽出精度の評価には，オントロジーに収められた用語を含む文章群から，人手でそれら用語を抽出した正解データが必要となる．今回対象とするオントロジーの中では，HPO のみ正解データが公開されている(HPO gold standard)[Groza 2015]．そこで，HPO gold standard(228 件のアブストラクトを含む)を用いて，各システムの処理速度と，適合率・再現率・F 値を比較した．

表 2 に HPO gold standard を用いた処理時間の比較結果を示す(**マシン環境**: Intel(R) Xeon(R) CPU E5-2697E@2.6GHz, 64G of RAM, CentOS release 7.3，**実行パラメータ**: デフォルト).

システム名	処理時間(sec)
PubDictionaries	554.6
MetaMap	351.0
NCBO Annotator	206.0
ConceptMapper	4.3

表 2. 情報抽出システムの処理時間比較

表 3 に HPO gold standard を用いた適合率, 再現率, F 値の比較結果を示す(**実行パラメータ**:デフォルト)

システム名	適合率	再現率	F 値
PubDictionaries	0.47	0.54	0.50
MetaMap	0.51	0.61	0.56
NCBO Annotator	0.54	0.47	0.51
ConceptMapper	0.52	0.51	0.52

表 3. 情報抽出システムの適合率・再現率・F 値

適合率, 再現率, F 値を比較した結果, 適合率は NCBO Annotator が最も高く, 再現率および F 値は MetaMap が最も高くなった. しかしながら処理時間は, ConceptMapper とその他のシステムとの間に大きな差があり, 2 番目に処理時間が短い NCBO Annotator と比較しても約 50 倍の差がある. 再現率と F 値が最も高い MetaMap を採用した場合, 1 つのオントロジーを用いて 100 万件のアブストラクトから情報を抽出するのに約 17.8 日を要するため, 実用的ではない(ConceptMapper は約 0.2 日). そのため, F 値が 2 番目に高く, 処理速度が実用的な ConceptMapper を採用し, 各症例報告から表 1 のオントロジーを用いて情報を抽出し, それらをアノテーションとして付与した.

2.3 類似症例報告検索手法

患者の症状・病態を元に, 各症例報告との類似度を計算する手法について以下に述べる. 各手法は, 類似度の降順で症例報告のリストを出力する.

(1) HPO を用いた手法 (以降, HPO-Sim と呼ぶ)

1つのオントロジーを用いた類似度計算手法は数多く開発されているが, その中でも Gene Ontology[GeneOntology 2017]を対象に開発された類似度計算手法である simGIC[Pesquita 2007]は, ライフサイエンスの分野で広く利用されている[Buske 2015]. この手法は, 予め概念毎の Information Content(*IC*)を計算しておく(式 2). これは, 式 1 で計算される概念 *c* およびその下位概念(is-a 関係)がコーパス中に出現する頻度(*annotc*)と, 全概念の総出現頻度(*annot*)を元に計算される確率 *P(c)*を利用する.

$$P(c) = \frac{|annotc|}{|annot|} \quad (1)$$

$$IC(c) = -logP(c) \quad (2)$$

式 3 で, simGIC による症例 *P* と症例報告 *Q* の類似度 *Sim(P,Q)* を計算する. *Concepts(P)*は症例 *P* に割り当てられた全ての概念 *c* とそれらの全ての上位概念(is-a 関係)のセットを表す.

$$Sim(P,Q) = \frac{\sum_{c \in Concepts(P) \cap Concepts(Q)} IC(c)}{\sum_{c \in Concepts(P) \cup Concepts(Q)} IC(c)} \quad (3)$$

(2) 複数オントロジーを用いた手法(以降, Multi-Sim と呼ぶ)

今回, 表 1 で示した複数のオントロジーを用いてアノテーションを付与し, それら全てのアノテーションを用いて式 4 [溝口 2008]で類似度を計算した. *R(P,Q)*は, *n* 個のオントロジーを用いた場合の症例 *P* と症例報告 *Q* の類似度となる. $Sim_i(P,Q)$は *i* 番目のオントロジーによって導き出された *P* と *Q* の simGIC による類似度である. w_iは *i* 番目のオントロジーによる類似度への重みを示す. 今回, FMA と MP の重みを 0.1 とし, その他を 0.2 とした.

$$R(P,Q) = \sum_{i=1}^{n} w_i * Sim_i(P,Q) \quad (4)$$

(3) Bag-of-Words を用いた手法 (以降, BoW-Sim と呼ぶ)

Bag-of-Words を用いる場合, 症例報告全体で出現した単語数を V とし, 各症例報告を V 次元のベクトルで表現する. ベクトルの各要素の値は, 該当する単語の TF-IDF[Zhang 2011]とした. 症例 P と症例報告 Q の類似度は, それぞれのベクトルのコサイン類似度とする.

(4) Doc2Vec を用いた手法 (以降, Doc2Vec-Sim と呼ぶ)

Mikolov らは文章の分散表現(ベクトル)を獲得する手法 PV-DM を提案し, Doc2Vec モジュールとしてその手法を実装した[Mikolov 2014]. この手法は, 任意の長さの文章に ID を付与し, その ID から生成される文章ベクトルと, 文章に含まれる単語から生成される単語ベクトルを同じベクトル空間に配置する. 設定した文脈(連続する単語)に続く単語を正確に予測できるように, ニューラル・ネットワークを用いて学習用の文章全体を学習し, 最終的に単語ベクトルと文章ベクトルを得る. 全症例報告を学習用データとし, 各症例報告のベクトルを得る. 症例 P と症例報告 Q の類似度は, それぞれのベクトルのコサイン類似度とする.

3. 類似症例報告検索手法の評価

ClinVar を用いて遺伝子変異と疾患が共通する症例報告セットを抽出し, 評価用データセットを作成した. その評価用データセットを用いて, 類似症例報告検索手法の精度を比較した.

3.1 評価用データセット作成

ClinVar は, 医療において重要な遺伝子変異と表現型(疾患)との関係性を取り扱うデータベースであり, データを FTP サイト(ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/)から取得できる. 今回, ClinVarFullRelease_2016-10.xml.gz を取得し, 205,738 件の遺伝子変異・表現型レコード(以降, VP レコードと呼ぶ)を抽出した. 希少疾患に関する VP レコードに限定するため, 表現型(疾患)に Orphanet へのリンクがある 57,857 件に絞り込み, さらに, Clinical Significance が"Pathogenic"または"Likely pathogenic"である 36,667 件に絞り込んだ. 各 VP レコードには, その遺伝子変異と表現型(疾患)の情報元である論文が紐づけられている. そこで, 2 件以上の症例報告が紐づけられている 806 件の VP レコードを取得した.

各 VP レコードの症例報告セットについて, 1 つをクエリとしてその他を類似症例報告と定義し, 2,011 件の「クエリ症例報告−類似症例報告」のセットを取得し, 評価用データセットとした.

3.2 評価方法と評価結果

各手法で, クエリの症例報告と症例報告全件との類似度を計算し, 類似度の降順で症例報告のリストを取得し, クエリと対になる類似症例報告の順位を求めた. 類似症例報告が複数存在する場合は, 最も高い順位を採用する. X 位以内のセットの数

が 2,011 件全体に占める割合を RecallX と定義し，各手法の RecallX を比較した(表 4). その結果，Recall10 と Recall20 では Doc2Vec-Sim が最も高い割合となり，その他は Multi-Sim が最も高い割合となった.

手法	Recall 10	Recall 20	Recall 50	Recall 100	Recall 1000
HPO-Sim	9.1% (184)	12.6% (254)	17.9% (360)	21.8% (439)	37.9% (763)
Multi-Sim	24.9% (501)	33.4% (671)	45.0% (904)	51.7% (1,039)	71.5% (1,437)
BoW-Sim	13.9% (280)	19.3% (389)	26.7% (537)	32.6% (656)	55.6% (1,118)
Doc2Vec-Sim	28.1% (566)	34.1% (686)	42.0% (844)	47.3% (952)	66.5% (1,337)

表 4. 類似症例報告検索手法の RecallX 比較

4. おわりに

Doc2Vec-Sim はクエリが文章であることを前提とした手法である. しかし，必ずしもクエリが文章とは限らず，ユーザがキーワードを入力する場合もある. クエリがキーワードの場合でも Multi-Sim は適用可能であり，その場合は Doc2Vec-Sim を除くと，Recall10 と Recall20 においも Multi-Sim が最も高い割合となる. また，既存の症例検索システムは類似症例検索に HPO のみを用いているが，全ての RecallX において Multi-Sim の割合が HPO-Sim と比べて約 2〜3 倍高く，HPO 以外のオントロジーも類似症例検索に有効であることを示唆している. さらに，類似文章検索に用いられる Bag-of-Words や Doc2Vec を組み込んだ手法と比較しても，一部を除いて Multi-Sim の RecallX の割合が高く，オントロジーを用いた類義語への対応および意味的検索が類似症例検索に有効であることを確認した. 今後は Multi-Sim を組み込んだ症例報告検索システムを，実際の臨床研究や未診断患者の診断に適用したい.

しかし，改善すべき課題も存在する. アノテーションを付与するオントロジーを任意に 6 つ選択したが，他にも生物医学に関するオントロジーは数多く存在し，本手法に最適なオントロジーを自動で選択する方法を考える必要がある. また，類似度を計算する際の重みも任意の値を割り当てたが，ベイズ最適化などのアルゴリズムを用いて，最適な重みを探索する必要がある.

5. 謝辞

本研究は MEXT 科研費 221S0002 の助成を受けたものです.

参考文献

[Orphanet 2017] Orphanet: http://www.orpha.net/consor/cgi-bin/index.html

[Mutarelli 2014] Mutarelli, Margherita, et al: A community-based resource for automatic exome variant-calling and annotation in Mendelian disorders. BMC Genomics, 15. 3 (2014): 1.

[Sawyer 2016] Sawyer, S. L., et al: Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clinical genetics, 89.3 (2016): 275-284.

[EURORDIS 2007] Survey of the delay in diagnosis for 8 rare diseases in Europe (EurordisCare2). Fact sheet EurordisCare 2.

[UDP 2017] UDP: https://www.genome.gov/27544402/the-undiagnosed-diseases-program/

[IRUD 2017] IRUD: http://www.amed.go.jp/program/IRUD/

[Laurel 2016] Willig, Laurel, K., et al: Whole-genome sequencing for identification of Mendelian disorders in critically ill infants: a retrospective analysis of diagnostic and clinical findings. The Lancet Respiratory Medicine, 3.5 (2015): 377-387.

[Leslie 2014] Biesecker, Leslie G., and Robert C. Green: Diagnostic clinical genome and exome sequencing. New England Journal of Medicine, 370.25 (2014): 2418-2425.APA.

[Philippakis 2015] Philippakis, Anthony, A., et al: The Matchmaker Exchange : A Platform for Rare Disease Gene Discovery. Human Mutation, 36. 10 (2015): 915-921.

[Sebastian 2017] Köhler, Sebastian, et al: The Human Phenotype Ontology in 2017. Nucleic Acids Research (2016): gkw1039.

[Carey 2010] John, C. Carey: The Importance of Case Reports in Advancing Scientific Knowledge of Rare Diseases. Adv Exp Med Biol, 686 (2010): 77-86.

[ClinVar 2017] ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

[PubMed 2017] PubMed: http://www.ncbi.nlm.nih.gov/pubmed

[BioPortal 2017] BioPortal: http://bioportal.bioontology.org

[Groza 2015] Groza, T., et al: Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database, bav005 (2015): 1-13.

[Taboada 2014] Taboada, M., et al: Automated semantic annotation of rare disease cases: a case study. Database, (2014): bau045.

[PubDictionaries 2017] PubDictionaries: http://pubdictionaries.org

[GeneOntology 2017] Gene Ontology Consortium: http://geneontology.org

[Pesquita 2007] Pesquita, C., et al: Evaluating go-based semantic similarity measures. InProceedings of 10thAnnualBio-OntologiesMeeting, 37. 40. (2007).

[Buske 2015] Buske, J., et al: PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. Human Mutation, 36. 10 (2015): 931-940.

[溝口 2008] 溝口祐美子: オントロジーを用いた文書間類似度計算手法. 人工知能と知識処理, 108. 119 (2008): 87-92.

[Zhang 2011] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang: A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38.3 (2011): 2758-2765.

[Mikolov 2014] Le, Quoc, V., and Tomas, Mikolov: Distributed Representations of Sentences and Documents. ICML, Vol. 14. (2014).