疾患プロファイル作成のための症状名抽出

野中真生*，奥村貴史†，建石由佳*，谷田和章‡，辻井潤一‡§

*工学院大学大学院工学系研究科情報学専攻

†国立保健医療科学院

‡東京大学大学院理工学系研究科情報科学専攻

§School of Computer Science, University of Manchester/National Centre of Text Mining

１．はじめに

医療が進歩した現代でも，治療方法が確立されていない難病が存在する．難病の情報を自動的に収集するために，患者にとって有益な情報を蓄えた診断支援システムを構築したいという需要がある[1]．この診断支援システムには疾患名と症状表現のデータが必要なため，入力された疾患に関する文書を解析して症状表現を抽出し，疾患プロファイルを作成する．本研究では，Metamap[1][2]を用いて文書を解析し，疾患に関する情報の抽出を試みた．その手法と結果について報告する．

２．手法

本研究では，疾患に関する文書から症状表現を抽出するために Metamap を用いた．Metamap は文書を分割して，Metathesaurus[2]というシソーラスを参照して，[Disease or Syndrome]などの生物医学概念タグを単語に対応付ける．例えば，“lumbar socliosis”という文字を Metamap で解析すると“lumbar socliosis”を“lumbar”と“soliosis”に分割し，それぞれに[Body Location]と[Anatomical Abnormality]という生物医学概念タグを対応付ける（図１）．本研究では，単語に対応付けられた生物医学概念タグの種類によって症状表現の判定を行った．また，連続する単語に付与されたタグの組み合わせにも着目した．連続する単語に量的概念や体の一部を表すタグが連続して付与されると疾患に関する情報である可能性があるためである．特定のタグの組み合わせが表れた場合，それを疾患に関する情報と判定する．例えば，連続する二つの単語に[Quantitative Concept]と[Body Part]のタグが付与された場合，その連続した単語は疾患に関する情報として抽出する．本研究では，Metamap が付与する生物医学概念タグを６つのグループに分類した．それぞれ，ACTION グループ，BODY グループ，FUNCTION グループ，QUALIFIER グループ，SYMPTOM グループ，ETC グループである．各グループに含まれる生物医学概念タグを表１にまとめた．ETC グループには表１に掲載されていない生物医学概念タグが含まれる．疾患に関する文書から抽出するタグの組み合わせは表２にまとめた．

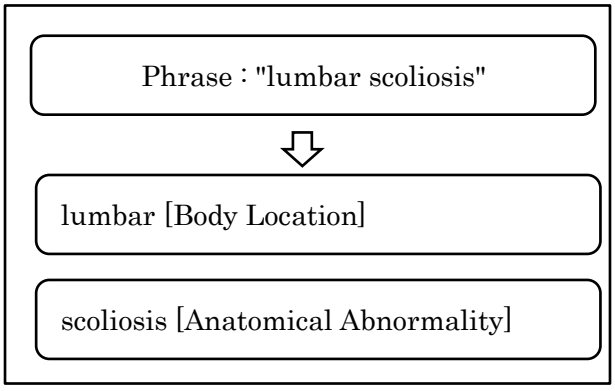

図１　Metamap の処理

[1] http://metamap.nlm.nih.gov/
[2] http://www.nlm.nih.gov/research/umls/

ACTION グループ
[Behavior] [Daily or Recreational Activity] [Environmental Effect of Humans] [Human-caused Phenomenon or Process] [Individual Behavior] [Social Behavior] [Therapeutic or Preventive Procedure]
BODY グループ
[Body Location or Region] [Body Part] [Body Space or Junction] [Body System] [Cell Component] [Embryonic Structure] [Fully Formed Anatomical Structure] [Organ, or Organ Component] [Physical Object] [Tissue]
FUNCTION グループ
[Biologic Function] [Cell Function] [Chemical Viewed Functionally] [Functional Concept] [Genetic Function] [Organ or Tissue Function] [Organism Function] [Pathologic Function] [Physiologic Function]
QUALIFIER グループ
[Biomedical or Dental Material] [Organism Attribute] [Qualitative Concept] [Quantitative Concept] [Spatial Concept] [Temporal Concept]
SYMPTOM グループ
[Acquired Abnormality] [Anatomical Abnormality] [Cell or Molecular Dysfunction] [Congenital Abnormality] [Disease or Syndrome] [Finding] [Injury or Poisoning] [Mental or Behavioral Dysfunction] [Sign or Symptom] [Virus]

表１　生物医学概念タググループ一覧

症状表現と判定できるタグの組み合わせ
{QUALIFIER}{FUNCTION}{BODY}
{QUALIFIER}{BODY}

表２　抽出するタグの組み合わせ一覧

本研究では，疾患に関する文書から，Metamap が対応付けるタグを利用して，疾患に関する情報の抽出を行った．情報抽出の流れを図２に示す．

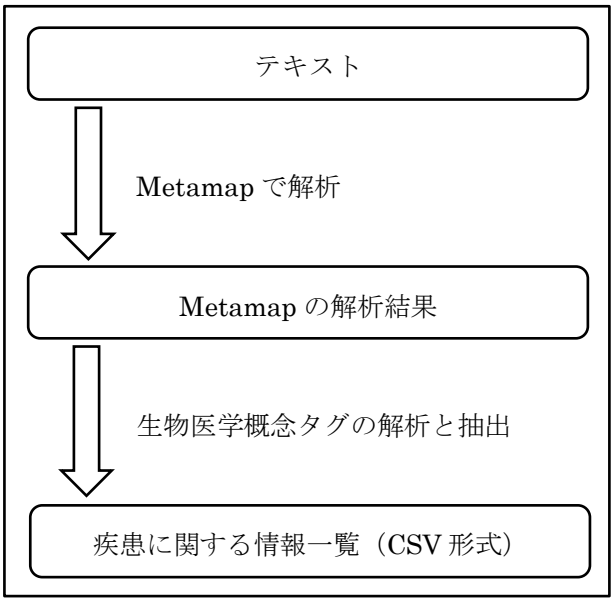

図２　情報抽出の流れ

まず，疾患に関する文書を Metamap で解析する．Metamap で解析された単語には生物医学概念タグが付与される．次に，単語に付与された生物医学概念タグが SYMPTOM グループに属するタグと一致するか判定する．また，連続するタグが表２の組み合わせであるか判定する．一致した場合，疾患に関する情報として抽出する．最後に，抽出した疾患に関する情報を CSV 形式でファイルに書き出して終了する．この手法を評価するために F 値を測定した．測定には，疾患に関する文書として OMIM[3]の臨床特徴（Clinical Features）を用いた．OMIM はオンラインで公開されている遺

[3] http://www.ncbi.nlm.nih.gov/omim

伝性疾患データベースである．OMIM に収録されている疾患には臨床特徴という疾患に関する情報が記述されている．また，OMIM の臨床特徴に人手で疾患に関する情報に正解タグを付与したデータを 20 文書用意した．本研究では正解タグが抽出した情報に含まれていれば正解とした（図３）．

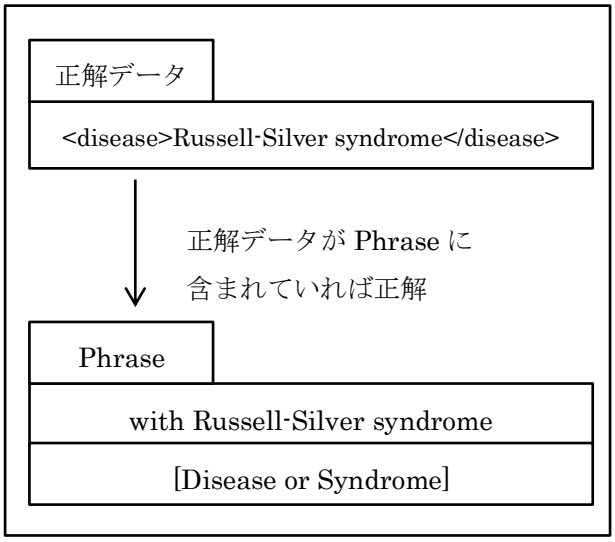

図３ 正解の判定

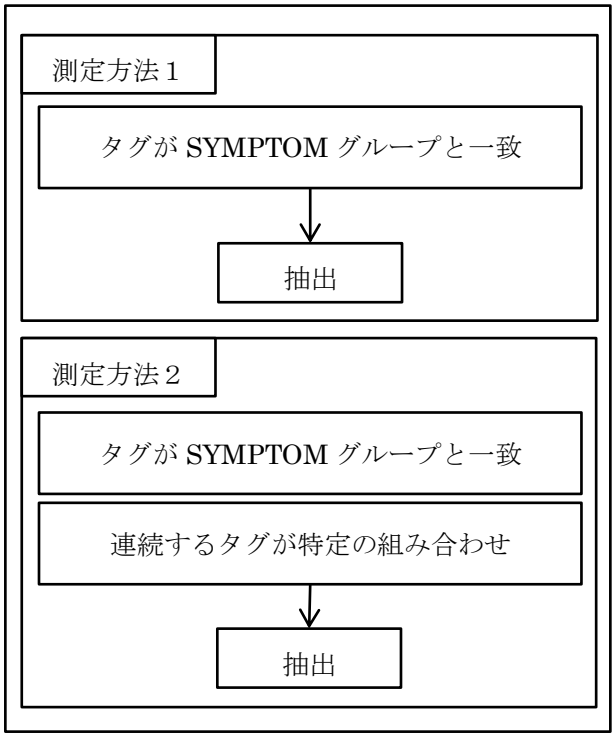

図４ 測定方法

測定は２種類の方法で行った（図４）．測定方法１は SYMPTOM グループに属するタグに一致したものを抽出する．測定方法２は連続するタグの組み合わせも利用する方法である．２種類の方法で測定した理由は，連続するタグの組み合わせが有効であるかを評価するためである．

３．結果

OMIM の臨床特徴から Metamap を用いて疾患に関する情報を抽出し，20 文書の正解データで評価した．その結果を表３に示す．表３に現れる C は正解文書に含まれる正解タグの総数である．N は本手法で抽出した疾患に関する情報の総数である．R は抽出した疾患に関する情報のうち正解であると判定されたデータの総数である．また，適合率（Precision），再現率（Recall），F 値（F-measure）は以下の式で算出される．

・$\text{Precision} = \frac{R}{N}$

・$\text{Recall} = \frac{R}{C}$

・$\text{F} - \text{measure} = \frac{2R}{N + C}$

	完全一致		図3の基準	
	測定1	測定2	測定1	測定2
最大適合率	33.3%	30.0%	75.0%	75.0%
最大再現率	23.1%	23.1%	83.3%	83.3%
最大Ｆ値	27.3%	26.1%	70.6%	70.6%
最小適合率	4.0%	5.6%	27.8%	27.8%
最小再現率	2.9%	5.9%	22.2%	25.9%
最小Ｆ値	3.4%	6.3%	25.0%	28.6%
平均適合率	11.8%	11.8%	49.5%	49.2%
平均再現率	14.2%	14.8%	61.3%	63.4%
平均Ｆ値	12.6%	12.8%	53.3%	54.2%

表３ 結果

４．考察

タグの組み合わせを用いた場合と用いなかった場合ではF値の平均に0.9%の違いが見られた．タグの組み合わせを用いた方がF値の平均が上昇したものの，上昇値は 0.9%と少ない．タグの組み合わせを用いても上昇値が少ないのは，Metamap が付与する１つのフレーズ[4](Phrase)の中にタグの組み合わせがないといけないとしたことが原因であると考えられる。そのため，タグの組み合わせを用いても F 値が 0.9%しか上昇しなかった. 20 文書から抽出した疾患に関する情報の合計 1824 個のうち，タグの組み合わせによって抽出されたデータは 78 個であった．さらに，その78個のうち正解であったデータは35個である．タグの組み合わせを用いてF値を上昇させるには，異なるフレーズの生物医学概念タグも抽出する必要がある．また，正解データと抽出した疾患に関する情報が完全一致する割合は約 13%であった．本手法の判定方法を用いることで54.2%までF値を上昇させることができた．

５．おわりに

本研究では，文書から疾患に関する情報の抽出を行った．疾患に関する文書を単語に分割し，生物医学概念タグを付与するツールとしてMetamap を用いた．疾患に関する文書として遺伝性疾患データベース OMIM の臨床特徴（Clinical Features）を利用した．20 文書の正解文書で測定した結果，タグの組み合わせを用いない場合，F 値の平均が 53.3%となり，タグの組み合わせを用いた場合，54.2%となった．タグの組み合わせを用いた場合の方が, F 値の平均が 0.9%上昇したが，正解文書の 20 文書と少なく，上昇値が小さいことから誤差の範囲内といえる．今後の課題として，異なるフレーズにまたがる生物医学概念タグとの組み合わせも考慮して測定を行う必要がある．また，正解文書を増やすことも重要である．

[4] Metamap は文章をフレーズ（Phrase）という単位で分割している．フレーズには１つの文の主語や目的語などが格納される．Metamap はフレーズ中の単語のうちシソーラスに一致した単語にだけ生物医学概念タグを付与する．１つのフレーズに付与される生物医学概念タグの数は１個である割合が高く．２個以上のタグがフレーズに付与されることは少ない．

参考文献

[1] 奥村貴史：臨床研究における症例登録と診断支援システム― 臨床医と患者の支援を通じた症例登録の促進に関する試論―，保健医療科学 Vol.59 No.3 pp. 212－217, 2010.

[2] Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In Proceedings of American Medical Informatics 2001 Annual Symposium, pp. 17–21, 2001.