ポスター
# ポスター16
# 用語定義・標準化とその利用
2017年11月23日(木) 09:15 〜 10:05 L会場（ポスター会場2） (12F ホワイエ)

## [4-L-1-PP16-1] 分散意味表現を利用したUMLSの概念と日本語の医学用語間のマッピングの試み

木村 映善, 蒲生 祥子, 石原 謙 （愛媛大学医学部）

統合医学用語システム(UMLS: Unified Medical Language System)は各情報源から電子化された生物医学情報を統合するシステムの開発支援を意図して設計された、メタシソーラス、語義ネットワーク、情報源マップ、SPECIALIST辞書よりなる知識ソースである。以来、 Watsonに代表されるように医療情報を機械学習に取り込むための有力なツールとして活用されている。我が国でも UMLSの邦訳が試みられているが、その成果は継続性のある形で公開されていない。 UMLSに収録されている概念の数は膨大であり、かつ継続的にアップデートしている。また翻訳は単なる表記上のマッチングのみならず、ドメイン固有の利用状況も踏まえる必要があるため、ドメインエキスパートの参加が不可欠である。日本語の医学用語を UMLSの概念に自動的にマッピングする手法を開発し、工数削減に貢献することを目標とした。各種ソースから英語-日本語のペアを抽出し、英和辞書を構築する。英和辞書に収録した日本語の用語は形態素解析の辞書にも取り込む。電子カルテの診療禄を形態素解析し、その結果をニューラルネットワークにかけて各単語の分散意味表現としての多次元ベクトルを獲得する。UMLSの概念に割り当てられた英語の代表的表記にマッチする英単語に対応する日本語の単語群を英和辞書から抽出する。その単語群を分散意味表現に置き換え、アウトライヤー処理し、重心となるベクトルを算出する。さらにそのベクトルに最も距離が近い分散表現を持つ単語を当該概念の「訳語」の候補として、アウトライヤー処理にて除外されなかった単語を当該概念の「同義語」とし、かつ同義語として提示したもののうち、編集距離が一定距離内にあるものを「異表記」候補として提示した。サンプリングした概念を検証した結果、機械的に英和の単語ベースでマッチングするよりも精度の高いマッピングが確認された。

# 分散意味表現を利用したUMLSの概念と日本語の医学用語間のマッピングの試み

木村映善[*1]、蒲生祥子[*1]、石原謙[*1]
*1 愛媛大学大学院医学系研究科博士課程医学専攻 社会・健康領域医療情報学講座

# Mapping between UMLS concepts and Japanese medical terms using distributed semantic representation

Eizen Kimura[*1] , Gamo Shoko[*1] , Ishihara Ken[*1]
*1 Dept. Medical Informatics of Medical School of Ehime Univ.

The Unified Medical Language System (UMLS) is the knowledge source to support a development of a system that retrieves and integrates electronic biomedical information from each data source. We aimed to develop a method to automatically map Japanese medical terms to UMLS concepts to contribute to the reduction of a workload of translation. Extract English-Japanese pairs from various sources and build English-Japanese dictionary. Japanese terms in the English-Japanese dictionary are also included in the morphological analysis dictionary. We applied the result of morphological analysis on the electronic medical record to the neural network to acquire the multidimensional vector as the distributed semantic representation of each word. A Japanese word group corresponding to an English word matching the representative expression of English assigned to the concept of UMLS is extracted from the English-Japanese dictionary. The words in the word group are replaced with its distributed meaning representation. Outlier processing is performed to exclude of misplaced translations. The vector serving as the centre of gravity of these vectors is calculated. The word having its vector closest in the distance to the vector of the centre of gravity is regarded as candidates for "translation word" of the concept. As a result of verifying with the sampled concept, a mapping with higher precision was confirmed than mechanically matching it on a word base of an English-Japanese dictionary.

Keywords: UMLS, fasttext, distributed representation, terminology.

## 1.目的

統合医学用語システム(UMLS: Unified Medical Language System)(1)は各情報源から電子化された生物医学情報を統合するシステムの開発支援を意図して設計された、メタシソーラス、語義ネットワーク、情報源マップ、SPECIALIST 辞書よりなる知識ソースである。以来、IBM 社 Watson に代表されるように、医療情報を機械学習に取り込むための情報処理基盤として貢献している。我が国でも先行研究においてUMLSの邦訳が試みられている(2, 3)が、継続的な取組みについて確認できていない。UMLS に収録されている概念の数は膨大であり、かつ継続的にアップデートされている。用語集の翻訳は単なる表記上のマッチングのみならず、その用語が使われている「コンテキスト」に照らして相応しい意味を持つ単語に訳する必要がある。その「コンテキスト」とはドメイン固有の利用状況であるため、ドメイン固有の事情に詳しいドメインエキスパートの参加が不可欠である。相当の人件費、工数が必要なことが翻訳作業の継続性を困難なものにしていることが考えられる。日本語の医学用語を UMLS の概念に自動的にマッピングする手法を開発し、工数削減に貢献することを目標とした。

## 2.背景

### 2.1 用語が指し示す概念と翻訳

語彙集の翻訳をする時に様々な課題があるが、本研究が対象としている課題を説明する。図 1 に提示した「青い丸」は、ある翻訳となる言語における用語が表している概念であるとする。(概念が定義する内容は円で表せるものではないが、あくまでも概念間の関係性を図示するための便宜として使用している。)この用語を別の言語(黄色で表現している)の用語に「翻訳」する場合、必ずしも概念が厳密に対応している用語が見つからないものである。A は最も概念が重なるもので、第一義の訳語として扱われるべきものである。B は概念にある程度の重なりがあり、同義語あるいは類義語に分類される。C、D は元の概念より狭義の概念を取り扱うもので、より具体的になる分、単語の対応だけではなく、その単語が使われている文脈から判断して適切な訳語を選択する必要がある。E はもともとの用語と関係がない訳語(翻訳・マッチングミス)や、異なる分野で使われている訳語が当該分野では適切ではないことがある。つまり元の言語では同じ表記であるが、翻訳先では分野によって捉え方が異なり、異なる意味・表記になるものがある。例えば、"system"は「体制、体系、組織網、系統、器官、身体、五体、全身、(情報)システム」などの様々な訳語が考えられる。用語が利用される文脈に応じて適切な訳語を採用する必要がある。

以上のように、ある言語の用語集を翻訳する場合は、各用語が指し示す概念と概念の広がり(広義・狭義)の関係性、そして翻訳先の言語の用語が持つ概念の関係性をも踏まえて適切な翻訳をする必要があるため、用語が使用されるドメインにおけるドメイン固有の概念、使われ方について熟知したドメインエキスパートの参画が必要となる理由である。

### 2.2 UMLS の階層化された概念と用語に関する仮説

UMLS は Metathesaurus で概念間の関係の定義や概念を

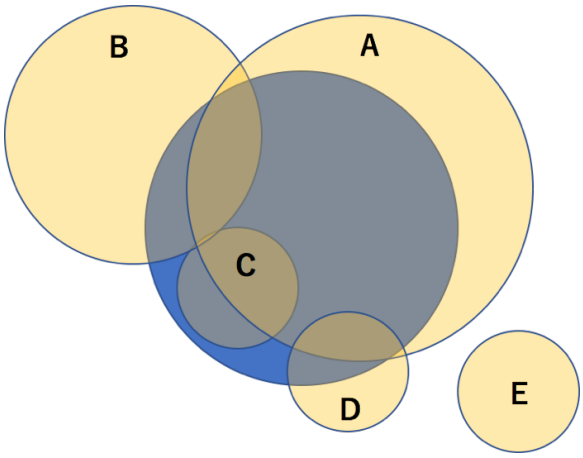


図 1 概念と用語の関係

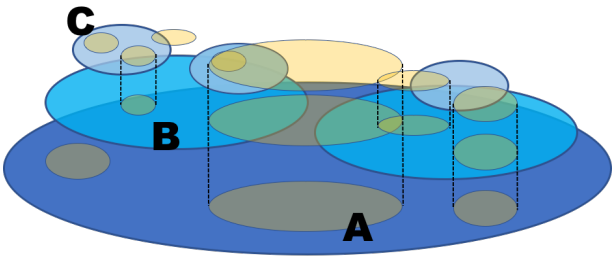


図 2 階層化された概念群と用語の関係

記述する用語とのつながりを定義している。以下に概念の階層と用語の関係を図示した図を提示する(図 2)。青系統で塗られている楕円は概念を表している。面積が大きいものから小さくなる(上方向に移動)につれて、より抽象な概念(上位概念)から具象的な概念(下位概念)となる。黄色で塗られている楕円は、その楕円が置かれている概念に対応する用語である。用語は複数レベルの概念で同時に使われることがある(図 2-A)。一方、より具体的な概念で初めてつかわれるケースもある(図 2-B、C)。用語の翻訳にあたり、可及的に具体的な概念にのみ対応する用語の候補を抽出できればいいが、現実的には図 2-A、B のように当該概念より上位の概念に対する訳語も候補として出現しうる。

ある概念に関して、抽象的なレベルから具体的なレベルまで順に具体化していく 3 つの概念 $C_1,C_2,C_3$($C_1$ が最も抽象的、$C_3$が最も具体的)があるとする。それぞれの概念に対する翻訳された用語(w)として、$C_1=\{w_1,w_2,w_3\}$,$C_2=\{w_1,w_4\}$,$C_3=\{w_4,w_5,w_6\}$が候補として抽出されるとする。より具体的な概念においては、上位概念に既に候補として出現したものは、より抽象的な概念の説明に使われるべき用語として候補から外す。すなわち、$C_3-(C_2 \cup C_1)$の処理を施して、残った候補が $C_3$ の翻訳された用語とする。一般化すると、$C_n$の翻訳用語は、$C_n$から$C_1 \ldots C_{n-1}$までの和集合を引いたものとなる。

## 2.3 翻訳タスクの定義

UMLS の用語集の翻訳を機械化する手法について仮説を立てる。インターネットやデータが提供されている英和辞典は、専門分野の英和辞典でない限り特定分野に依らず網羅的に訳語を収録していると考えられる。図 1 の例示で言えば、対象とならない「E」の様な訳語や誤訳も相当数混じっている状態である。字面から単純に翻訳候補として選択した単語群から、医学分野の諸概念を収録した UMLS の概念群に対する訳語としてふさわしい単語を抽出するには、候補の単語が「医学」に関連したコンテキストで使用されていることがわかっている単語群の中から選択することが正解に近づく方法であると考える。具体的には次のような仮定をおく。

単語を医学分野の用語として使用している文章の集合体として電子カルテのテキストを利用する(図 3)。ある概念によく対応するものは電子カルテにおける意味分散表現としてのベクトル群で中心的な位置をとる(図 3 において、A→II,III の関係)。周辺・部分的にマッピングされる用語は、同様に電子カルテから獲得した意味分散表現でも周辺的なベクトル群にマッピングされると仮定する(図 3 において、C→I,III、B→IV,V の関係)。当該概念から離れてマッピングされた用語は、電子カルテ上でも同様に他ベクトルから離れたベクトルとして位置する(図 3 において、D→VII の関係)。逆に、電子カルテから得られたベクトル群に対して外れ値の検出タスクを実施し、外れ値としてみなされたベクトルに対応する単語は、大元の概念の訳語としては相応しくないと判断する。

すなわち、(1)医学分野の用語として使われているという前提に妥当性を与えられるような語彙ソースの選択、(2)単語間の分散表現の獲得、(3)ある用語の翻訳としての単語群の獲得(辞書検索)、(4)単語間のクラスタリング、外れ値処理、(5)外れ値を除いた単語群を候補として提示、(6)上記までの処理を各概念に実施し、2.2 節で紹介した概念の階層を配慮した上で絞り込んだ単語を最終候補として提示する。さらに候補単語群のベクトル重心から、距離が近いものを同義語、異表記候補として提示する、という一連の処理によって用語集の精度の高い翻訳が可能となるという仮説を立てる。

本研究では、(1)-(5)に関してフィージビリティ評価を行う。

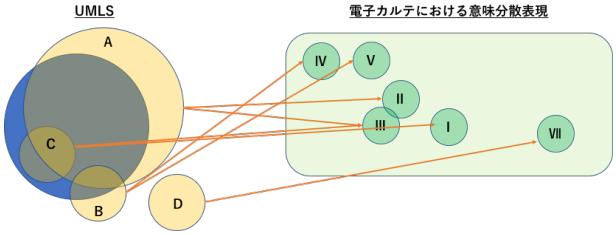


図 3 UMLS の概念の訳語と電子カルテから抽出した意味分散表現間の対応

## 2.4 意味分散表現

単語の分散表現(distributed representation)は単語の分散状況を実数値ベクトルで表す試みであり、word2vec(4)が高速に分散表現を獲得する実装として知られている。以下、word2vec の Skip-Gram(5)の実装を例にまとめる。Skip-Gram は 2 層のニューラルネットワーク、隠れ層は 1 つのみで構成されている。それぞれの層のユニットは隣接する層と全結合している。ある文章(英語の場合は単語の集合、日本語の場合は分かち書き後の結果集合)の中に「ある単語」と、「ある単語」の周辺に共起する「単語群」を抽出する。共起する単語は、前後ウィンドウサイズ(n 単語分)が対象となる。教師あり学習として、「ある単語」は入力層に、「ある単語」に共起していた「単語群」は出力層に配置する。語彙を直接ニューラルネットワークに投入できないため、文章を構成する全ての語彙をリストアップしたボキャブラリを構築し、それぞれの語彙を one-hot ベクトルとして表現する。全文章中のユニークな語彙が n 個とすると one-hot ベクトルは n×1 次元のベクトルとして表現

される。学習される単語ベクトルを m 次元とすると、隠れ層は n×m 次元の行列で表現される。教師あり学習の過程で隠れ層の n×m 次元の行列が学習されていくが、これが単語の分散表現に対応する単語ベクトルとなる。さらに隠れ層から出力された 1×m の単語ベクトルは、隠れ層～出力層間の重み付けが掛けられた後に出力層に出力される。学習の過程で、実際に共起した単語間の内積が大きくなるように重みを調整していく。

最終的に得られた単語ベクトルの特徴量には加法構成性があることが知られており、概念的に類似性のある単語の探索や、概念の足し算らしきものが出来ることが確認されている(6)。本研究では、このような分散表現を表すベクトルを利用して訳語間の概念的な近接性の算出に利用する。

## 3.方法

### 3.1 医療情報の収集

診療禄に記載されている文章および、その文章を構成している単語群は、医学分野の概念を表すために使用されているという仮説を立てた。愛媛大学医学部附属病院の電子カルテシステムに記載された SOAP 文書から収集し、Comejisyo(7)と mecab-ipadic-NEologd(8)をユーザ辞書として指定した medab(9)で分かち書きを生成する。Comejisyo は医学用語に特化した辞書である。Mecab-ipadic-NEologd は、未知の単語、形態素が原因の解析誤りによる分かち書き性能の低下を改善すべく、新規語彙・固有表現をインターネットから蒐集し、基礎語彙数の数的拡充に注力した辞書である。

### 3.2 英和辞書の作成

英和辞書として、JST 科学技術用語日英対訳辞書(2015 年版)、日外 25 万語医学用語大辞典、英次郎 v.138、ライフサイエンス辞書、研究社 医学英和辞典第 2 版から、英語と日本語の訳語を抽出して MySQL サーバに英和の検索データベースを構築した。英語のエントリーは米国医学図書館から提供されている SPECIALT Lexicon ツール(10)を利用して正規化した。日本語のエントリーは、neologd で利用されている正規化ルーチンを利用した。

### 3.3 UMLS の概念と英和辞書のマッチング

UMLS における概念間の階層構造 (MRREL テーブルにて定義)の中で末端に位置 (当概念において子となる概念関係（CHD: Child）がないもの) する概念であり、かつ親関係(PAR: Parent)、兄弟関係(SIB: Sibling)の関係性の定義数が多い上位群から無作為に検証候補となる概念を選択する。

その検証候補としての概念に紐付けられている用語が正規化されたもの(NSTR: normalized string)を MRXNS_ENG テーブルから抽出し、前述で構築した英和辞書の正規化された英語のエントリーとマッチングし、当該エントリーに対応する日本語を訳語の候補として抽出する。このような抽出をした理由は、末端の方がより具体的な概念であり、特定の用語と対応づけやすいことと、親・兄弟の関係性が多いことにより同義語から多くの翻訳候補が出現し、後述する外れ値の算出評価によいサンプルとなると判断したためである。

### 3.4 医療情報中の単語の分散表現の獲得

解析環境として Mac OS X Sierra、Intel Xeon E5 3Ghz 8-Core x2 64GB RAM、Anaconda 4.3.0,python の機械学習ライブラリ sklearn(11)を使用した。

単語の分散表現は、word2vec(4)がよく知られているが、未知語に対しても文字列から意味ベクトルを推定できる subword information(12)という理論を取り入れた fasttext(13)を利用して学習した。医学用語は複数の単語の組合せからなる用語も多いため、mecab が使用する辞書に収容されていない用語についても意味を適切に推論できることを期待してのことである。

自然言語処理に特化した機械学習ライブラリ gensim(14)から fasttext を呼び出して、前項で処理した分かち書き済みファイルを 300 次元の skipgram モデル、コンテキストウィンドウサイズは 8、12 スレッドの並列処理で学習させた。

### 3.5 外れ値の算出

3.3 節で訳語の候補として抽出された単語について、fasttext で抽出した分散表現としてのベクトルを抽出する。候補単語数が n 個とすると、300 次元のベクトルが n 個ある集合として定義される。この集合の中から外れ値を検出して、外れ値に対応する単語を訳語の候補から除外する。

外れ値検出アルゴリズムとして sklearn で実装されている One-Class SVM(15) 、 elliptic envelope(16) 、 Isolation Forest(17)¶、Local Outlier Factor(18)を使用した。各分類器の初期パラメータは sklern で設定されているデフォルト値を適用した。

なお、今回は 2.3 節で定義した翻訳タスクのうち、(1)～(5)までのステップのみであるため、概念上の抽象化・具体化の方向での広がりは外れ値として検出しなくても許容した。例:「ケロイド座瘡」に対して「項部ケロイド座瘡」。このケースでは「項部」という条件を加えて、より具体的な概念となっている。但し、「項部」のみ抽出された場合、本来の「ケロイド座瘡」とはつながりがないので、外れ値として分類する。

## 4.結果

電子カルテシステムから抽出されたデータは、延べ 10325079 行、4016152166 単語より構成される分かち書きデータとなった。全体でユニークな単語数は 371684 となった。

ランダムに抽出された 10 個の UMLS の概念を表 1 に示す。各々の概念に対して提示された訳語候補としての単語の総数は 218 個であった。候補各々について、分類器が外れ値として検出したものについて、人間が判断して結果的に適切な訳語ではなかった場合は TP: True Positive、適切な訳語であった場合は FP: False Positive に分類した。4 種の外れ値検出手法を評価した結果を表 2 に提示する。全体的に One Class SVM が優位な結果を出しているが、全体的にスコアが低い Isolation Forest は Precision のみ高スコアであった。

サンプルとして、MRXNS_ENG テーブルから抽出された、C0019202 に対応する NSTR 一覧を提示する(表 2)。図 4 は、候補単語群の分布を t-SNE を用いて次元圧縮・可視化したものである。各単語の座標に 4 分割（左上:One-Class SVM、右上:Elliptic Envelope、右下:Local Outlier Factor、左下:Isolation Forest）されている正方形をプロットした。外れ値として検出した場合は赤色の正方形で表示している。

表 1 本実験で無作為抽出された UMLS の概念

| |
|---|
| C0001145 acne cheloidalis |
| C0009069 blood clot evaluation retraction time |
| C0019202 cerebral disorder pseudosclerosis |
| C0021171 ambiguous disorder incontinentia pigmenti syndrome |
| C0032805 cardiotomy post syndrome |
| C0038505 angiomatosis calcification cerebral meningofacial syndrome |
| C0043119 adult age premature syndrome |
| C0085292 baby stiff syndrome |
| C0175693 diagnosis russell silver syndrome |
| C0696113 assay ferritin serum |

表 3 C0019202 に対応する NSTR(MRXNS_ENG)

| | |
|---|---|
| cerebral disorder pseudosclerosis | diagnosis disease wilson |
| cerebral pseudosclerosis | dis hepatoneurologic wilson |
| chorea gowers | dis kinnier wilson |
| chorea gower | dis wilson |
| copper disease storage | disease disorder wilson |
| d1 hepatoneurologic wilson | disease hepato neurologic wilson |
| d1 kinnier wilson | disease kinnier wilson |
| d1 wilson | disease or syndrome wilson |
| degeneration disease find hepatolenticular | disease strumpell westphal |
| | disease wd wilson |
| degeneration disease finding hepatolenticular | disease wilson |
| | disorder strumpell syndrome westphal |
| degeneration hepato lenticular | familial hepatitis |
| degeneration hepatocerebral | pseudosclerosis westphal |
| degeneration hepatolenticular syndrome | pseudosclerosis |
| | strumpell syndrome westphal |
| degeneration hepatolenticular | strumpell westphal |
| degeneration lenticular progressive | wd |
| degeneration neurohepatic | wnd |

# 5.考察

## 5.1 分類器の選択

Isolation Forest の仮定は「正常値は数が多く相互に近似していること、外れ値は数が少なくて相互に異なること」としている。外れ値となるような単語が少なく、外れ値に対応する単語が外の単語から離散している場合は Isolation Forest による適合率は高まることが考えられる。しかし、再現率が低いので、他の分類器と皮下くして外れ値に対して緩やかな基準で取りこぼしている可能性がある。英和の自動マッピングにあたっては、本来訳語の候補とすべきものを外れ値として扱わないような安全側に倒れた判断をし、最終的な判断を人間に仰ぐことが望ましいと思われる。すなわち、今後の実装を進めるにあたり、全般的に高評価であった One Class SVM を主な分類器として採用し、One Class SVM で外れ値として検出されても、Isolation Forest による判定と対立すれば、外れ値としない処理を加えることを検討する。一連のマッピングの後、2.4 節で述べた「(6)概念の階層を配慮して絞り込む処理」によって、外れ値の境界上の評価にあったものがふるい落とされる可能性があるためである。

## 5.2 意味分散表現の解釈

図 4 を俯瞰すると、核レンズ変性症の表記揺れ・同義語について、比較的正確に正常範囲とみなしているが、「同值性」「関連性」「水曜日」「手関節離断」等、訳語として認めるべきではないものも混同している。

表 2 分類器による外れ値検出の評価

| | OneClassSVM | Elliptic Envelope | Local Outlier Factor | Isolation Forest |
|---|---|---|---|---|
| Accuracy | **0.665** | 0.656 | 0.628 | 0.573 |
| Precision | 0.228 | 0.215 | 0.139 | **0.247** |
| Recall | **0.600** | 0.567 | 0.458 | 0.352 |
| F-measure | **0.330** | 0.312 | 0.214 | 0.290 |

これは略語を中心とした、他の概念でも使われている用語を概念の英語の表記へのマッピングの過程で引きずり込んでしまっているものである。表 3 に掲示しているように「wd」「wnd」は Wilson Disease の略語として使われるであろうが、他の意味にも捉えられる。それだけであれば、意味が違うものが入るので、外の正当な単語のベクトルから距離が離れそうなものであるが、そうなっていない(例:図 3 で「水曜日」は外れ値として分類されていない)。電子カルテの記載の中に肝レンズ核変性症に関する記述と、水曜日の記述が近接している箇所があり、意味分散表現上は近い距離として評価されためと思われる。約 40 億単語より構成されている医療情報であっても、このような意図しない意味分散表現上の近接性を排除するのには不十分な可能性がある。

## 5.3 分かち書きと sub word information 理論

医学用語は複数の単語の組合せからなる用語も多いため、構成要素となる単語単位で分かち書きしないように、固有表現を多く網羅するべく Comejisyo と mecab-ipadic-NEologd をユーザ辞書として利用した。fasttext における sub word information 手法が日本語についてどのように作用しているか未確認である。固有表現を網羅することを試みるよりも、意味のある範囲で最小限の長さを持つ単語のみで分かち書きするようにした方が、より実際の概念の関係・操作の結果を提供する意味分散表現を獲得できる可能性も考えられる。


図 4 C0019202 の訳語候補の可視化

# 6.結語

本研究は UMLS の日本語訳を可能な限り自動化することを目標とし、以下の仮説のもとに検証した。

電子カルテのデータは医療分野に関して精度の高い意味分散表現を獲得することに貢献する。不適切な訳語は他の訳語のベクトルから距離があるので、外れ値検出によって除外できる。単階層でのマッピング結果に対して、UMLS 上の階層関係を通してさらに絞り込みをかけることで精度の高い

マッピングが実現できる可能性がある。

外れ値としては One Class SVM がもっともよい成績を示した。外れ値として除外できなかったものは、主に略語による同表記異義語が混入することと、大元のデータソースの情報量が足りないことによる十分な意味分散表現を獲得できなかったことによる可能性がある。

今回得られた知見をもとにマッピングプロセスの改善と、上下概念のマッピング結果を加味することにより、マッピングの精度を高めていきたい。

## 7. COI、謝辞

本研究に関して申告すべき COI はありません。愛媛大学臨床研究倫理審査委員会「1707001 号 高品位な知識抽出を実現する三階層オントロジーフレームワークの開発」の承認のもとデータ収集・分析を実施した。

## 参考文献

1. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl_1):D267-D70.
2. 田中 昌昭. 医学用語の UMLS へのマッピングの試み. 2000;20 (Suppl. 2):920-1.
3. 脊山 洋右, 開原 成允, 野添 篤毅, 小野木 雄三, 篠原 恒樹, 鈴木 博道. UMLS と連携した日本語医学用語シソーラスの開発実験. 医療情報学連合大会論文集. 2004;24th:1202-3.
4. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:14023722. 2014.
5. Guthrie D, Allison B, Liu W, Guthrie L, Wilks Y, editors. A closer look at skip-gram modelling. Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006); 2006: sn.
6. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, editors. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems; 2013.
7. 相良かおる. ComeJisyo の紹介と医療情報に含まれる誤字調査. 情報知識学会誌. 2014;24(2):204-9.
8. Toshinori Sato TH, Okumura M, editors. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese)2017: The Association for Natural Language Processing.
9. Kudo T. Mecab: Yet another part-of-speech and morphological analyzer. http://mecab/ sourceforge net/. 2005.
10. McCray AT, Srinivasan S, Browne AC, editors. Lexical methods for managing variation in biomedical terminologies. Proceedings of the Annual Symposium on Computer Application in Medical Care; 1994: American Medical Informatics Association.
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825-30.
12. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. arXiv preprint arXiv:160704606. 2016.
13. Joulin A, Grave E, Bojanowski P, Douze M, Jégou H, Mikolov T. FastText. zip: Compressing text classification models. arXiv preprint arXiv:161203651. 2016.
14. Rehurek R, Sojka P, editors. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010: Citeseer.
15. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural computation. 2001;13(7):1443-71.
16. Rousseeuw PJ, Driessen KV. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999;41(3):212-23.
17. Liu FT, Ting KM, Zhou Z-H, editors. Isolation forest. Data Mining, 2008 ICDM'08 Eighth IEEE International Conference on; 2008: IEEE.
18. Breunig MM, Kriegel H-P, Ng RT, Sander J, editors. LOF: identifying density-based local outliers. ACM sigmod record; 2000: ACM.