原著

テキストマイニングを用いた教科書からの専門用語の抽出 —放射線技術学領域における用語集の更新に向けて—

辻真太郎¹ 谷川原綾子² 福田晋久³ 谷川琢海² 川真田実⁴
西本尚樹⁵ 島井健一郎⁶ 星野修平⁷ 小笠原克彦¹

¹ 北海道大学大学院保健科学研究所

² 北海道科学大学保健医療学部

³ 北海道医療センター放射線科

⁴ 大阪国際がんセンター放射線診断・IVR科

⁵ 香川大学医学部附属病院臨床研究支援センター(現 北海道大学病院臨床研究開発センター)

⁶ 千葉大学医学部附属病院企画情報部

⁷ 群馬パース大学保健科学部

論文受付

2017年4月12日

論文受理

2017年9月13日

Code No. 1000

緒言

日本学術会議の活動に協力する「協力学術研究団

体」(平成28年3月31日現在, 2008学会)は, 日本学術会議, 日本学術協力財団, 科学技術振興機構が共同

Applying Text-mining to Extracting Technical Terms from Textbooks —Toward Updating the Terminology in the Field of Radiology Technology—

Shintaro Tsuji,¹ Ayako Yagahara,² Akihisa Fukuda,³ Takumi Tanikawa,² Minoru Kawamata,⁴
Naoki Nishimoto,⁵ Kenichiro Shimai,⁶ Syuhei Hoshino,⁷ and Katsuhiko Ogasawara^{1*}

¹ Faculty of Health Sciences, Hokkaido University

² Faculty of Health Sciences, Hokkaido University of Science

³ Department of Radiological Technology, Hokkaido Medical Center

⁴ Department of Diagnostic and Interventional Radiology, Osaka International Cancer Institute

⁵ Clinical Research Support Center, Kagawa University Hospital

(Current address: Clinical Research and Medical Innovation Center, Hokkaido University Hospital)

⁶ Division of Medical Informatics and Management, Chiba University Hospital

⁷ School of Radiological Sciences, Faculty of Health Science, Gunma Paz University

Received April 12, 2017; Revision accepted September 13, 2017

Code No. 1000

Summary

Although terminology requires continuous consideration of recorded technical terms, extracting these terms manually is difficult, because the number of recorded terms is constantly increasing. Text-mining acquires information from numerous documents, and is capable of extracting technical terms. The purpose of this study is to extract candidate terms using text-mining toward updating the terminology of Japanese society of radiological technology (JSRT). First, the subjects for this study were textbooks published by the JSRT, and morphological analysis was conducted, which is an analysis to break the books up into meaningful words. Additionally, index terms of textbooks were extracted. Second, we observed overlaps between the JSRT technical terms and the terms obtained from the morphological analysis and the indexes of textbooks and the extracted terms were absent in the JSRT terminology. The overlap was 53.6% (3090/5770 terms). The terms, "imaging technology for magnetic resonance" and "information and system in radiological technology" were missing from the JSRT terminology. From these results, it was estimated that half number of the JSRT technical terms were changing with time. This study demonstrated that text mining showed the differences between old and new technical terms.

Key words: terminology, technical term, morphological analysis, compound terms, text mining

*Proceeding author

運営する「学会名鑑」というデータベースに登録されている¹⁾。学会名鑑に登録されている学会は、農学や医学・歯学、薬学などを対象領域とする「生命科学」、工学や理学などを対象領域とする「理学・工学」、政治学や経済学、心理学など対象領域とする「人文・社会科学」の三つに分類(以下、大分類)されており、これらの三つの大分類の下位には、更に詳細な分類(以下、小分類)が行われている。筆者らの先行研究では、学会名鑑の小分類を対象とした用語集の構築状況の調査によって、放射線技術学を含む小分類「臨床医学」が他の小分類と比較して最も多くの用語集が構築されていることが明らかとなった(167学会中20学会)²⁾。また、用語集が公開されている媒体に関する調査では、紙媒体から portable document format (PDF) を主とした電子媒体での公開に移行するとともに、今後は web を介した用語の検索・参照サービスの提供が広く普及する可能性があることも明らかとなった。更には、近年提供される用語集は電子媒体化が進み、ページ数などの制約がなくなったことにより用語の参照サービスと合わせて日本外科学会用語集(38495語)といった大規模な用語の提供が可能となっている。

科学技術において専門用語集の構築が必要とされる理由には、専門用語集が学際的なコミュニケーションを促進することが挙げられる³⁾。すなわち専門用語集を構築する利点とは、(1)専門用語の背景にある概念の整理が可能であること、(2)英語を主とした他言語への翻訳が可能であること、(3)情報の圧縮、すなわち用語の概念が共通認識として得られている場合、詳細の説明を省略できることが可能なこと、(4)情報の検索が可能であること、が挙げられている。(2)の利点については、筆者らの先行研究が示したように、既に多くの用語集が英語の用語に対する日本語の対訳を中心に行っている²⁾。また(4)の利点は、前項にて述べた近年の用語集の電子化とそれに伴う用語検索サービスの提供により重視されている。加えて、(3)と(4)については、用語集の設計と用語の選定を通じて実現されるが、その工程において「専門用語の範囲の決定方法と候補となる語の選定方法」「掲載する用語の統一方法と用語の意味の記述方法」に関する二つの検討事項が先行研究から指摘されている³⁻⁷⁾。

専門用語集に掲載すべき用語の範囲の決定では、一般的な語と専門用語の領域の線引きをどこで行うかという問題が生じる。原則として用語集の構築の指針によってその範囲の設定を事前に行うことで解決される。用語集の構築の指針を検討する方法としては、United Nation Scientific and Cultural Organization

(UNESCO)によって提示されている Guideline for Terminology Policies⁷⁾を参考にすることができる。このガイドラインは、学術的な専門用語集だけではなく、国家レベル、地域(数カ国)レベル、国際レベルで構築される用語集のデザイン、指針の計画と改善、用語集の管理について基本的な指針を示している。専門用語集の指針が決定した後の用語選定のアプローチ方法としては、トップダウンに概念を絞り込み、用語を選定する方法と、専門分野の教科書の目次に基づいた方法が提案されている⁴⁾。一方、教科書などに収載されている用語一つ一つを専門用語の候補として検討し、ボトムアップに概念を検討していく方法も提案されているが⁶⁾、事例に基づいた先行研究は少ない。上述の語の検討方法で共通の問題となるのは、専門家による用語の判定前に行われる候補のピックアップであり、近年の用語集の大規模化により手作業による負担が大きくなっていることである。

前述の検討課題に加え、医学用語の専門用語集の構築においては、医学用語の特徴として、取り扱う用語数が膨大で、複合語が多く、その組み合わせが無数にあるという問題が指摘されている⁸⁾。筆者らの先行研究においても、「医学」関連の用語集の収録語数が、他の分野と比較して多いことを示した²⁾。例えば、日本の学会は、「生命科学」「理学・工学」「人文・社会科学」の三つの分野に分けられるが¹⁾、「生命科学」で最大の語数を持つ用語集は日本外科学会(38495語)であり、「理学・工学」に属する日本原子力学会(7545語)、「人文・社会科学」の日本図学会(2987語)と比較して多い。対象とする学術領域における専門用語の抽出には教科書の目次や索引を用いる方法が提案されているが⁴⁾、複合語の抽出となると医療における複合語の特徴⁸⁾からもその組み合わせが多いことから用手的に選出することは不可能である。

一方、筆者らは、日本放射線技術学会が提供している放射線技術学用語集⁹⁾(Japanese Society of Radiological Technology: JSRT 用語集)の更新に向けた用語集の分析を行ってきた¹⁰⁾。JSRT 用語集は、現在紙媒体での提供が行われており、2003年を最後に用語の見直しが行われていない。学術団体として専門領域の用語を定期的に更新する必要がある一方で、用語集の更新には語の網羅性を加味しつつ用手的に語を取捨選択する必要がある、人的な労力を考えると容易ではない。

したがって本研究では、テキストマイニング技術を用いた専門用語の候補の抽出を目的として、日本放射線技術学会が監修する教科書である放射線技術学シリーズを対象とした形態素解析を行い、得られた語と

Table 1 Details of the textbooks

Titles of textbooks	Edition/Print	Year of issue
MRI evaluation and analysis	1st/1st	August 2012
Radiation biology	2nd/5th	January 2014
Radiation safety management	2nd/4th	September 2013
Physics and clinical use of X-ray CT	2nd/3rd	March 2013
Emergency radiology	1st/1st	March 2011
Image quality measurement of digital radiography	1st/3rd	December 2013
Information and systems in radiological technology	1st/4th	August 2013
CT image measurement	1st/5th	December 2013
Imaging technology for radiographs	1st/5th	April 2013
Radiation therapy technology	1st/9th	October 2013
Radiation physics	1st/4th	January 2000
Radiation measurement	2nd/1st	November 2013
Nuclear medicine technology	2nd/3rd	January 2009
Imaging technology for magnetic resonance	2nd/9th	February 2014
Radiochemistry	2nd/6th	December 2012

JSRT 用語集の用語間の重複を調査して専門用語の候補を明らかにした。

1. 対象と方法

1-1 対象

JSRT 用語集に追加する語の候補(以下、候補語)を抽出する対象として、日本放射線技術学会が監修を行っている書籍「放射線技術学シリーズ」(全15冊、教科書)を用いた。これらの教科書は、分析時ににおいて入手可能な最新版を用いた(入手時期:2015年5月)。使用した教科書の詳細を Table 1 に示す。また、JSRT 用語集は2003年の補遺編を含むすべての用語(5770語)を対象とした。

1-2 方法

本研究では JSRT 用語集に追加する候補語を、用語集が対象とするすべての教科書の索引を利用する方法と、すべての教科書の本文を対象とした形態素解析によって抽出する方法を併用して候補語を抽出した。以下に候補語の抽出方法と候補語の分類について説明する。

1-2-1 専門用語の候補となる語の定義と分類

初めに、JSRT 用語集に掲載されている語の集合を $Terms_{Terminology}$ とし、すべての教科書の本文の形態素解析と索引から得られる語の集合を $Terms_{Textbook}$ 、両者の重複した語を除いた語の集合を $Terms_{Candidates}$ とすると、以下の式が成り立つ。

$$Terms_{Textbook} - (Terms_{Terminology} \cap Terms_{Textbook}) = Terms_{Candidates} \quad \cdots \text{式(1)}$$

$$\text{ただし、} (Terms_{Terminology} \cap Terms_{Textbook}) \neq \phi$$

$Terms_{Textbook}$: すべての教科書から得られた語の集合

$Terms_{Terminology}$: JSRT 用語集の語の集合

$Terms_{Candidates}$: すべての教科書から得られた候補語の集合

次に、すべての教科書から得られた候補語の集合 $Terms_{Candidates}$ に対し、すべての教科書の索引から得られた語の集合を $Terms_{Index}$ 、すべての教科書本文の形態素解析によって得られた語(形態素)で名詞を構成する語の集合を候補語 $Terms_{Morpheme}$ とすると式(2)が成り立つ。

$$Terms_{Candidates} = Terms_{Index} \cup Terms_{Morpheme} \quad \cdots \text{式(2)}$$

$$\text{ただし、} (Terms_{Index} \cap Terms_{Morpheme}) = \phi$$

更に、 $Terms_{Index}$ については、複数の教科書で使用されている索引語の集合として $Terms_{Index-Overlapped}$ 、単一の教科書で索引として使用されている索引語の集合を $Terms_{Index-Non overlapped}$ とすると、式(3)が成り立つ。

$$Terms_{Index} = Terms_{Index-Overlapped} \cup Terms_{Index-Non overlapped} \quad \cdots \text{式(3)}$$

$$\text{ただし、} (Terms_{Index-Overlapped} \cap Terms_{Index-Non overlapped}) = \phi$$

$Terms_{Index}$: すべての教科書の索引部分から得られた候補語

$Terms_{Index-Overlapped}$: 複数の教科書で使用されている索引語

$Terms_{Index-Non overlapped}$: 単一の教科書で使用されている索引語

最後に、 $Terms_{Morpheme}$ は、これらの候補語の中で、

Fig. 1 Category of candidates of technical terms.

索引語を含む候補語の集合を $Terms_{Morpheme-Index}$ 、索引語を含まない候補語の集合を $Terms_{Morpheme-Non index}$ とすると式(4)が成り立つ.

$$Terms_{Morpheme} = Terms_{Morpheme-Index} \cup Terms_{Morpheme-Non index} \quad \cdots \text{式(4)}$$

$$\text{ただし, } (Terms_{Morpheme-Index} \cap Terms_{Morpheme-Non index}) = \phi$$

以上の式(1)～(4)に基づいて求められる索引語 $Terms_{Candidates}$ の分類の概要を Fig. 1 に示す.

1-2-2 JSRT 用語集の候補語を抽出する方法

本項では JSRT 用語集の候補語を抽出する方法につ

いて述べる (Fig. 2). まず初めに, 教科書(全 15 冊)は, スキャナの optical character reader (OCR) を用いてテキストデータ化し, 診療放射線技師 9 人にて用手的に誤字脱字の修正を行った. 上記のテキストデータの誤字脱字の修正を担当した診療放射線技師については, 各技術学の領域に関して業務経験ないし研究の経験がある者が担当した. また, 落丁や乱丁については, デジタル化されたテキストデータと該当するすべての教科書のページを目視にて比較しながら文字の整合性の確認を行った.

続いて, すべての教科書の索引と本文の文書に対してそれぞれ別の処理を行った. 一つ目の処理は, 索引語に対する処理として, すべての教科書の索引語の間に重複している語の削除と, すべての教科書の中で複数回使用されている索引語の特定を行った. 二つ目の処理は, すべての教科書の本文に対する処理として, 文書を最小の品詞単位(形態素)に分割し, 品詞・活用形・読みなどを与える処理である形態素解析を行い, 教科書の本文から専門用語の候補語を抜き出すために形態素の中から名詞を抽出した. 形態素解析器には MeCab (ver. 0.996) を用いた. 放射線技術学領域で用

Fig. 2 Overview of method.

Table 2 Combination of compound terms

Types of compound terms	Patterns of compound terms
Terms _{Morpheme-Index}	
Terms _{Morpheme-Non index}	

○○○ shows index term. ●●● shows non index term (noun).

いられる用語は形態素解析器の辞書への登録数が少ないため⁹⁾、本研究では「教科書において索引語が重要な語として位置付けられている」という前提の下、形態素解析器 MeCab に標準的に搭載されている IPA 辞書に加え、前項で述べた索引語 (Terms_{Index}) を辞書として形態素解析器に搭載した。得られた形態素の中で、名詞句の複合語となる語でかつ、索引語を含むか否かの条件で複合語を二つに分類した (Table 2)。本研究における名詞句の定義は、「文書内にて名詞以外に囲まれた名詞句は複合語である」とした。以下、上記の定義を満たす名詞句を複合語と述べ、得られた複合語は、索引語を含むか否かで二つに分類 (Terms_{Morpheme-Index} と Terms_{Morpheme-Non index}) を行い、それぞれの候補語の語数を明らかにした。

1-2-3 分析に使用した教科書と候補語の評価

式(1)で示したとおり、用語集の語と本研究で抽出した候補語 (Terms_{Index-Overlapped}, Terms_{Index-Non overlapped}, Terms_{Morpheme-Index}, Terms_{Morpheme-Non index}) には重複が存在する可能性がある。仮に用語集の語と候補語が重複する割合を 100% とすると、本研究で使用したすべての教科書が JSRT 用語集の領域をすべて網羅している可能性が高いと考えられる。そこで本研究では、分析に使用した教科書の網羅性の指標として、JSRT 用語集と候補語の重複の割合 (Overlapping ratio) を算出した。定義式を以下に示す。

$$\text{Overlapping ratio}[\%] = \frac{\text{Terms}_{\text{candidates}} \cap \text{Terms}_{\text{Terminology}}}{\text{Terms}_{\text{Terminology}}} \times 100[\%]$$

..... 式(5)

JSRT 用語集の更新作業では、式(1)で求めた JSRT 用語集との重複した語以外は新規に取り扱う可能性のある候補語となる。そこで、候補語全体に対する用語集との重複した語 (Overlapping ratio) を差し引いた語数の割合を新規候補語の出現割合 (New candidate ratio) として以下の式で定義した。

$$\text{New candidate ratio}[\%] = (1 - \text{Overlapping ratio})[\%] \quad \cdots \text{式(6)}$$

式 (5) と式 (6) の指標を求めることで、Terms_{Index-Overlapped}, Terms_{Index-Non overlapped}, Terms_{Morpheme-Index}, Terms_{Morpheme-Non index} と Terms_{Terminology} との重複が明らかとなる。次にこれらの語が今回使用したどの教科書から抽出された語かを明らかにするために、あらかじめ上記の語には、掲載されていた教科書名を属性として保存し集計に利用した。また、特定の教科書における Overlapping ratio を Overlapping ratio_{Specific textbook} [%] として式(7)により算出した。更に、JSRT 用語集に対して今後特定の分野の専門用語の追加を検討するために特定の教科書における New candidate ratio_{Specific textbook} [%] として式(8)により算出した。

Overlapping ratio_{Specific textbook}

$$= \frac{(\text{Terms}_{\text{candidates}} \cap \text{Terms}_{\text{Terminology}})_{\text{Specific textbook}}}{\text{Terms}_{\text{Terminology}}} \times 100[\%]$$

..... 式 (7)

New candidate ratio_{Specific textbook}

$$= \frac{(1 - (\text{Terms}_{\text{candidates}} \cap \text{Terms}_{\text{Terminology}})_{\text{Specific textbook}})}{1 - (\text{Terms}_{\text{candidates}} \cap \text{Terms}_{\text{Terminology}})} \times 100[\%]$$

..... 式(8)

2. 結果

2-1 索引語および複合語の語数

教科書の索引語は、述べ語数で、8357 語、重複を除いた異なり語数は、7340 語となった。このうち、複数の教科書で使用されていた索引語 Terms_{Index-Overlapped} は、770 語、単一の教科書で使用されていた索引語 Terms_{Index-Non overlapped} は 6570 語であった。複数の教科書で使用されていた索引語の例を Table 3 に示す。また、すべての教科書の本文の形態素解析から、名詞および名詞句を形成する形態素が述べ 2755660 語得られた。これらの語は、意味をなさない記号を取り除くデータクリーニング後に、索引語を含む複合語 Terms_{Morpheme-Index} が述べ 29290 語、異なり語数 26510 語得られた。一方、索引語を含まない語 Terms_{Morpheme-Non index} は述べ 203292 語、異なり語数で 59605 語得られた (Table 4)。

2-2 分析で使用した教科書の評価

JSRT 用語集と本研究で得られた候補語と重複する語数と重複の割合 (Overlapping ratio) を求めた結果、用語集と候補語との重複は 53.6% であった (Fig. 3)。また、その内訳は、複数の教科書で使用されている索引語 Terms_{Index-Overlapped} との重複が 235 語 (4.1%)、単一の教科書で使用されている索引語

Table 3 Examples of Terms _{Index-Overlapped}

Frequency of technical terms	Radiochemistry	Imaging technology for radiographs	Imaging technology for magnetic resonance	MRI evaluation and analysis	Image quality measurement of digital radiography	Radiation biology	Radiation physics	Radiation therapy technology	Radiation measurement	Radiation safety management	Nuclear medicine technology	CT image measurement	Physics and clinical use of X-ray CT	Terms Index-Overlapped
7	-	-	-	-	-	-	○	○	○	○	○	○	○	線減弱係数
7	-	-	-	-	○	○	○	○	○	○	○	-	-	放射能
6	-	-	-	-	○	○	-	○	○	-	○	○	-	ポアソン分布
5	-	○	○	○	-	-	-	-	-	-	○	-	○	アーチファクト
5	○	-	-	-	-	-	○	-	○	○	○	-	-	PET
5	-	-	○	-	○	-	○	-	-	-	-	○	○	フーリエ変換
5	-	-	-	-	○	-	○	-	○	○	○	-	-	ベクレル
5	-	-	-	-	-	○	-	○	○	○	○	-	-	吸収線量
5	-	-	-	-	-	-	○	○	-	○	○	○	-	空間分解能
5	-	-	-	-	-	○	○	○	○	-	○	-	-	光電効果
5	-	-	-	-	-	○	-	-	-	-	○	○	○	散乱線
5	-	-	○	-	-	-	-	-	-	○	○	○	○	投影データ
5	-	-	-	-	-	-	○	○	-	○	-	○	-	半価層
5	-	-	-	-	-	-	-	-	-	○	○	○	○	標準偏差
5	-	-	-	-	-	○	○	○	○	○	-	-	-	放射損失
5	○	-	-	-	-	○	○	○	○	-	○	-	-	励起
Frequency of textbooks	2	1	3	1	8	7	9	9	10	8	13	8	6	

Table 4 Number of candidates of technical terms

Category of candidates	Number of terms	Total number of terms
Terms _{Index-Overlapped}	770	8357
Terms _{Index-Non overlapped}	6570	
Terms _{Morpheme-Index}	26510	32410
Terms _{Morpheme-Non index}	59605	205941
Total (Terms _{Candidates})	93455	246708

Terms_{Index-Non overlapped}との重複が 553 語 (9.6%), すべての教科書の本文の形態素解析から得られた語で索引を含む複合語 Terms_{Morpheme-Index}との重複が 1439 語 (24.9%), 索引語を含まない語 Terms_{Morpheme-Non index}との重複が 863 語 (15.0%)となった. 幾つかの例を

Table 5 に示す.

2-3 用語集に追記する候補語の評価

教科書の形態素解析で抽出した語および索引語のうち、用語集と重複しなかった新規候補語の出現率

Fig. 3 Overlapping ratio and breakdown of the candidates in the JSRT terminology.

Table 5 Examples of the overlapping terms between the JSRT terminology and the candidates of technical terms

Relationships between JSRT terminology and the categories of candidate	Examples of the terms
Terms Terminology $\cap$ Terms Index-Overlapped (N=235)	線減弱係数, アーチファクト, 放射能, フーリエ変換, ベクレル, 光電効果, 散乱線, 投影データ, 半価層, 放射損失, 筋起, β 線, エネルギーフルエンス, コンプトン効果, フィルタ補正逆投影法, レントゲン, 管電圧, 血液脳関門, 後方散乱, 質量減弱係数, 消滅放射線, 熱中性子, 濃度, 半導体検出器, 付加フィルタ
Terms Terminology $\cap$ Terms Index-Non overlapped (N=553)	アルゴリズム, アレイコイル, イオン再結合, イオン再結合補正, イニシエーション, イントラネット, インフロー効果, コーンウィナースベクトル, ウロキナーゼ, エコー時間, オージェ電子, エスケープピーク, エネルギー単位, エリアシング, オルソフィルム, コリメータ散乱係数, コーンビーム, コバルト校正定数
Terms Terminology $\cap$ Terms Morpheme-Index (N=1439)	連続パルス励起法, 励起パルス図, 臨床標的体積, 量子力学的軌道運動, 粒子線治療装置, 粒子フルエンス, 離散のフーリエ変換, 卵巣癌, 陽電子断層撮影, 陽子線, 預託占領, 有効視野, 門脈, 無停電電源装置, 密封小線源治療, 放射能汚染, 放射線技術学, 閉鎖孔, アクリル樹脂ファントム, QA プログラム, PKI 個人認証
Terms Terminology $\cap$ Terms Morpheme-Non index (N=863)	脾シンチグラム, 脛骨前縁, 嚥下運動, 量子力学的カロリメータ, 両側ラウエンシュタイン, 硫化コバルト沈殿法, 留置ステント, 輸血用血液照射装置, 薬事関係法規, カッピングアーチファクト, ガスフロー係数装置, ガイディングカテーテル, オーダ発行機能, ウェッジフィルタ角度, インビボ分子イメージング技術, イオン対生成

(New candidates ratio) は, 96.7%であった (Table 6). また, Fig. 1 の候補語の分類毎の出現率は, Terms Index-Overlapped が 69.5% (535 語), Terms Index-Non overlapped が 91.6% (6017 語), Terms Morpheme-Index が 94.6% (25071 語), Terms Morpheme-Non index が 98.6% (58742 語)であった.

2-4 得られた形態素が引用される教科書の割合

Terms Index-Overlapped における Overlapping ratio Specific textbook

Table 6 Number of new candidate terms and its ratio

Category	Number of new candidates (%)
Terms Index-Overlapped	535/770 (69.5)
Terms Index-Non overlapped	6017/6570 (91.6)
Terms Morpheme-Index	25071/26510 (94.6)
Terms Morpheme-Non index	58742/59605 (98.6)
Total (Terms Candidates)	90365/93455 (96.7)

Fig. 4 Overlapping ratio of Terms Specific textbook Index-Overlapped

Fig. 5 Overlapping ratio of Terms Specific textbook Morpheme-Non index

は、全体の 16.7% を放射線計測学が占めており、次いで放射線治療技術学 14.6%、放射線安全管理学 12.7% となった (Fig. 4)。また、Terms Morpheme-Non index における Overlapping ratio Specific textbook は、全体の 17.0% を放射線計測学が占めており、次いで放射線治療技術学

11.6%、放射線安全管理学が 10.4% であった (Fig. 5)。

一方、Terms Morpheme-Index における New candidate ratio Specific textbook は、全体の 14.6% を magnetic resonance imaging (MRI) 撮影技術学が占めており、次いで MRI の評価と解析が 9.1%、放射線計測学 8.7% と

なった(Fig. 6). また, Terms_{Morpheme-Non index} における New candidate ratio_{Specific textbook} は, 全体の 12.0% を放射線システム情報学が占めており, 次いで MRI 撮影技術学 10.1%, 放射線安全管理学 9.8% となった(Fig. 7).

3. 考 察

本研究は, 学術団体が運用している専門用語集の更新時における候補語を, 自然言語処理技術の一つである形態素解析を用いて大量のテキストデータから抽出する手法を提案した. 本研究の背景には医学における専門用語の特徴として「複合語の組み合わせの多さ」が指摘されており, これらを手動的に抽出することが困難であること, 更には近年の用語集の電子化に伴い用語集に掲載する語数が増加していることが挙げられる. 以下に, 使用した教科書, Overlapping ratio, New candidate ratio, 候補語の出典先から本研究の手法と結果の妥当性について検討を行う.

3-1 分析で用いた教科書と Overlapping ratio

JSRT 用語集と本研究で得られた候補語とが重複する割合(Overlapping ratio)から, 分析に使用した教科書 15 冊から得た候補語は, 用語集のおよそ 53.6% (3090 語) をカバーしていることが明らかになった(Fig. 3). 一方, 重複している語(3090 語)に占める複合語の割合は 74.5% (2302 語) となり, 索引語から得た候補語だけではなく, それらに関連した複合語も JSRT 用語集に含まれていることが示された. この事実は, 先行研究で指摘されている医療における複合語の多さを裏付けている結果となった. また, Overlapping ratio を左右する要因には, 下記の三つがあると考えられる.

一つは参考文献(本研究では教科書)の数とその領域の要因と考える. この理由は, 本研究が 15 冊の教科書を用いたのに対して, JSRT 用語集は 63 以上の教科書, 文献, 論文, 用語集, ガイドラインを参考に構築された. つまり, JSRT 用語集の構築に参照された文献の種類や数と, 今回使用した 15 冊の教科書という単一の参考文献の種類を比較すると今回使用した教科書の網羅性は低いと断言できないと考えられる. したがって, 今後は教科書以外の文献, 論文, 用語集, ガイドラインから候補語を抽出することで Overlapping ratio の向上が期待される.

次に二つ目の要因として, 用語集構築時の組織の体制が考えられる. 現在の JSRT 用語集は 1989 年に検討が始まり, 撮影, 画像, 核医学, 治療の四つのワーキンググループが検討作業を担当したとされてい

る¹⁰⁾. 本研究結果から, 式(8)の New candidate ratio_{Specific textbook} をみると「放射線システム情報学」が最も高い 12% であり, JSRT 用語集に「放射線システム情報学」の用語を追加する必要があることを示している. JSRT 用語集に不足している教科書の語数の割合である式(8)から「放射線システム情報学」が 12% と最も高くなった. 上記の四つのワーキンググループでは補いきれなかった領域の語が現在必要となることが Overlapping ratio が 50% 程度にとどまっている原因の一つと考える.

最後に三つ目の要因は, 時間の経過による用語の変化である. 放射線技術の発展とともに必要となる新たな知識が増加し, その一方, 利用しなくなった技術に関する語が教科書から消えることで用語集と教科書から得た候補語のギャップが生じている可能性が考えられる. 例えば, JSRT 用語集には, 放射線写真学で用いられる用語で「酢酸繊維素」「超微粒子現像法」「MQ 現像液」等の用語の記載があるが, 本研究で使用した教科書の本文・索引には出現しなかった. これは近年, computed radiography の普及で現像液を使用した X 線撮影を実施する施設が減少し, 更に 2004 年から診療放射線技師国家試験の科目からも外れたことにより上記の用語の重要性が低下したためと考えられる.

3-2 複合語から得た候補語と選定作業

本研究の結果から, 形態素解析によって得られた複合語には用語集に掲載されている用語が多く含まれており, 用語集の更新には複合語を効率的に抽出することが必要不可欠であることを明らかにした. しかし, Terms_{Index-Non overlapped}, Terms_{Morpheme-Index}, Terms_{Morpheme-Non index} のそれぞれ約 90% (90365 語) が新たに追加の可能性のある候補語となっており, 現在の用語集の規模(約 5800 語の掲載語)を考慮すると, その 15 倍の候補語を検討するのは現実的ではない. したがって, 本研究結果を活用した候補語の選定作業は, 索引語由来の候補語の検討を先に行い, 複合語に関しては手動的な検討に入る前に候補語の数を減らす前処理を行い, 手動的に検討するコストを低減させたのちに選定作業を行う必要がある. 具体的な方法としては, 複合語における共通の語を中心とした共起語(特定の語と一緒に使われる語)を求めて提示する方法がある. 今後は, 複合語の処理の問題⁸⁾をいかに効率的に処理するかが用語集の継続的な更新に必要であると考えられる.

3-3 複合語から得た候補語と引用する教科書

本研究では形態素解析によって得た語がどの教科書

Fig. 6 New candidate ratio Specific textbook of Terms Morpheme-Index.

Fig. 7 New candidate ratio Specific textbook of Terms Morpheme-Non index.

から抽出されたかを明らかにした。この分析から以下の二つの知見が得られたと考える。一つ目は、用語集と重複した候補語には、放射線計測学、放射線治療技術学、放射線安全管理学の教科書から抽出された語が多いことから、これらの分野の語が15年間の時間経過にもかかわらず放射線技術学用語の中で重要な位置を占めていることが考えられる。また、上記の教科書で使用されている語は複数の教科書に使用されていることから、放射線技術学における中心的な語であることが推測される。この推測を裏付ける事実として索引語の数が考えられる。すなわち、放射線計測学の索引語は767語、放射線治療技術学は515語、放射線安全管理学は666語であり、最も索引語の多いX線撮影技術学(1234語)と、それに次ぐMRI撮影技術学(873語)を比較すると索引語の多さによって放射線計測学、放射線治療技術学、放射線安全管理学の候補語が抽出されたことは考えにくい。二つ目は、用語集と重複しなかった候補語では、索引語を含む複合語 Terms Morpheme-Index として、MRI撮影技術学とMRIの評価と解析が上位となった。つまり、MRI関連の用語は索引語を中心とした複合語の組み合わせが多く存在していることが考えられる。更に索引語を含まない複合語 Terms Morpheme-Non index については、放射線情報システム学、MRIの評価と解析が上位となった。上記の教科書には索引語にはない複合語が多数出現しており、用語集が2002年を最後に更新が行われていない事実を考慮すると、近年の情報技術とMRIの目覚ましい発展とともに新しい用語が発生したと考えることができる。

3-4 掲載する用語の統一と用語の意味の記述

専門用語の選定には、類似した用語を統一し、「概念」として記述する必要がある。本研究で扱う「概念」とは、用語集が対象としている利用者が専門用語を参照した際に想起するイメージと定義する。理想的には、一概念に対して一つの専門用語を定義することができれば、その専門用語を用いた論文などの検索では漏れは生じることはない。しかし、現実には一つの概念に対して一つの専門用語を定義することは困難であり、複数の語が一つの概念に対応することを許容せざるを得ない。本研究で得られた候補語の中にも一つ概念を表すが、異なる表記をもつ語が散見された。例えば、「鉄線量計」と「フリッケ線量計」、「複合核モデル」と「複合核模型」、「点応答関数」と「点拡がり関数」と「点広がり関数」、「周波数エンコーディング」と「周波数エンコード」などが挙げられた。

一方、独断的に用語の定義を押し進めれば、利用者のコミュニケーションにおいて混乱状態を引き起こす可能性があり避けなければならないことが指摘されている³⁾。特に医学領域の用語集の構築に関しては、学会単位で行われることが多いが、複数の領域を含む用語集の場合、関係する領域の用語集との調和性を保ちつつ用語集の構築を行わなければならない。専門用語の科学的な分析を行う専門用語学では、上記の問題に対処するため、用語集にオントロジー技術を用いることが提案されている。オントロジーは、「対象とする世界の情報処理モデルを構築する人が、その世界をどのように眺めたか、つまりその世界に何が存在するか」を前提に可視化したものであると定義されている^{11~14)}。具体的には、オントロジーは、概念と概念間の関係を明らかにすることであり、一つ概念に対して「特定の条件」、つまり用語集の対象領域に限定した定義を行うことが可能である。実際に欧米では International Health Terminology Standards Development Organization(IHTSDO)による医学・医療全般をまかなう総合的な用語集として Systematized Nomenclature of Medicine-Clinical Terms(SNOMED-CT)¹⁵⁾の提供がオントロジー技術の利用を背景として行われている。放射線技術学領域の専門用語集においても、関連学会との調和を考慮しつつ本研究で示したような大規模な候補語に基づいた大規模な用語集の構築が行われることが予想され、先行研究でも述べたオントロジーの構築が必要となると考えられる²⁾。

3-5 展望と課題

本研究で用いた形態素解析により大量の専門用語の候補を抽出することが可能となり、合わせて出典先の教科書の情報(索引、教科書の領域)を活用することで、用語集の傾向を推測する手がかりを得ることが可能になると考えられる。本手法を用いることによって用語の作業にかかわる人員の作業の負担を低減させるとともに、語の統計的な分析に基づく用語集の構築に貢献することが可能となると考えられる。また、将来的には、上記の情報に加え、用語とその属性(読み、英語訳、略語など)と統合することでJSRT用語集をオントロジーとして構築することが可能となると考えられる²⁾。

4. 結語

本研究では、テキストマイニング技術を用いた専門用語の候補の抽出を目的として、日本放射線技術学会

が監修する教科書である放射線技術学シリーズを対象とした形態素解析を行った結果、抽出された 2755660 語の形態素とすべての教科書の索引語 8357 語のうち既存の用語集(5770 語)との重複は 53.6%であった。また、候補語は 90365 語となり、索引語を含む複合語の組み合わせの多い MRI の領域と、索引語にも掲載されない新たな語を多く含む医療情報と MRI 領域の

用語が不足していることが明らかとなった。

本研究は、公益社団法人日本放射線技術学会平成 26, 27 年度学術調査研究班の研究テーマ「放射線技術学分野における電子用語集の構築を目指した知識基盤の整備」として行った。

参考文献

- 1) 日本学術会議, 日本学術協力財団, 科学技術振興機構. 学会名鑑. <https://gakkai.jst.go.jp/gakkai/> (2017/2/23 確認).
- 2) 辻真太郎, 福田晋久, 谷川原綾子, 他. 放射線技術学領域における用語オントロジーの構築に向けた分析. 日放技学誌 2015; 71(3): 186-193.
- 3) Helmut F, Christian G, Christian G. 大島富士子訳. 学問としての用語学—ターミノロジー学に関するウィーン学派としての考察—, 情報管理 1982; 25(8): 659-670.
- 4) 長尾 真. 辞典形式での専門分野の知識の体系的構成法. 人工知能学会誌 1992; 7(2): 320-328.
- 5) 仲本秀四郎. 用語「情報」: ターミノロジー的考察. 情報の科と技 2002; 52(6): 339-342.
- 6) 守屋秀夫. 学術用語集 建築学編(増訂版)の編集からみた専門用語の問題点(〈特集〉ターミノロジー). 情報の科と技 1991; 41(4): 344-350.
- 7) United Nations Educational, Scientific and Cultural Organization. Guidelines for terminology policies. <http://unesdoc.unesco.org/images/0014/001407/140765e.pdf> (2017/2/23 確認).
- 8) Rector, AL. Clinical terminology: why is it so hard? Methods Inf Med 1999; 38(4-5): 239-252.
- 9) 日本放射線技術学会. 放射線技術学用語集. 放射線医療技術学叢書(5). 京都: 日本放射線技術学会, 1994.
- 10) 辻真太郎, 西本尚樹, 小笠原克彦. 形態素解析における放射線技術学分野の用語適用—診療放射線技師試験を対象とした未知語の調査. 日放技学誌 2008; 64(7): 791-794.
- 11) 来村徳信, 溝理一郎. オントロジー工学に基づく機能的知識体系化の枠組み. 人工知能学会誌 2002; 17(1): 61-72.
- 12) 溝口理一郎. オントロジー研究の基礎と応用. 人工知能学会誌 1999; 14(6): 1019-1032.
- 13) 溝口理一郎. 知の科学 オントロジー工学. 東京: オーム社, 2005.
- 14) 溝口理一郎, 来村徳信, 古崎晃司. オントロジー構築入門. 東京: オーム社, 2006.
- 15) SNOMED International. SNOMED CT. <http://www.ihtsdo.org/snomed-ct/> (2017/02/23 確認).

問合せ先

〒060-0812 札幌市北区北 12 条西 5 丁目
北海道大学大学院保健科学研究所 小笠原克彦