チャンキング及び体言判別を用いた専門用語の自動抽出手法 － 放射線技術学関連の医学用語抽出への応用 －

谷川原 綾子[*1], プタシンスキ ミハウ[*2], 辻 真太朗[*3], 上杉 正人[*4]
[*1] 北海道科学大学保健医療学部, [*2] 北見工業大学情報システム工学科,
[*3]Department of Digital Health Sciences, Mayo Clinic, [*4] 北海道情報大学医療情報学部

A Method for Automatic Extraction of Technical Terms using Chunking and Noun-phrase Discrimination - With Application to Extraction of Medical Terms Related to Field of Radiological Technology -

Ayako Yagahara[*1], Michal Ptaszynski[*2], Shintaro Tsuji[*3], Masahito Uesugi[*4]
[*1] Hokkaido University of Sciences [*2] Kitami Institute of Technology
[*3] Mayo Clinic, [*4] Hokkaido Information University

抄録：放射線技術学用語の整備は手作業による編集作業は多大な時間と人的資源を要するため，機械による支援が求められる．本研究では、放射線技術学に係る専門用語の自動抽出法について検討を行った．手法では放射線技術学に関する教科書から，N-gram にて単語抽出を行い，句読点等の不要な記号を削除した．次に，助詞の前に現れる体言を抽出し，専門用語候補を抽出した．候補語の調整として，一般的な日本語の語彙を削除した．その結果，1143 語が抽出された．そのうち 309 語が日本放射線技術学会の用語集に掲載されている語と一致した．用語集に含まれていない候補語の 834 語のうち 774 語が専門家 2 名以上により専門用語と判定された．本手法は専門用語の抽出に有用である可能性が示唆された．

キーワード 自然言語処理，形態素解析，用語抽出，放射線技術，N-gram

1. はじめに

放射線技術分野において用語集の定期的な整備は，学術的なコミュニケーションの促進や，自然言語処理の精度向上において，非常に重要なタスクである．しかし，手作業での整備は，多大な時間と人的資源を必要とするため，機械による支援が求められる．先行研究[1]では，放射線技術学用語集の更新に向けた新規専門用語の抽出を行った．教科書から，名詞が単独もしくは連続して出現した語を新規専門用語とした結果，約 10 万語が抽出されたが，放射線技術学用語ではない用語も多いことが問題であった．そこで本研究では，上記問題を解決するために，専門用語のみを特定するための手法について検討を行った．

2. 放射線技術学用語抽出法

1) 文書の準備

放射線技術に関する教科書として先行研究[1]と同様に，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，入手時期：2015 年 5 月)を使用した．教科書の本文をテキスト化し，1 つのファイルにまとめ Cabocha[2]を用いてチャンキングを行った．

2) N グラムの抽出とクリーニング

本研究では，チャンキングで得られた文節の N グラムを用いる．N の範囲を従来研究[3]にて提案された基準をもとに 1~7 で設定した．抽出後の N グラムには，専門用語抽出を防ぐ不要な句読点や，未閉じの括弧などが残されていることが多いため，句読点や記号を各 N グラムのチャンクから削除した．

3) 専門用語候補の抽出

2)で抽出された N グラムの中から専門用語候補(候補語)とその頻度を計数した．専門用語には体言として考えられるものは多い．体言とは，言語学において語形変化をしない語彙であり，日本語では名詞，代名詞が該当する．本研究では，体言を認識するために，日本語の体言への総合的アプローチ[4]を用い，助詞の前に現れた語彙を抽出した．また，用語には一般的に考える文字

(カタカナ, 漢字など)以外の記号も含まれていることはあるが, 今回は, Proof-of-concept としてひらがな, カタカナ, 漢字という 3 つの成分のどれか, あるいはその組み合わせから構成された用語のみを抽出した. さらに, 1 文字のみ, ひらがなのみ, 出現頻度が 1 回のみの語を削除した.

4) 候補語の調整

候補語の中から, 専門的なコンテキストだけで現れるフレーズのみを抽出するため, 残った候補語から一般的な日本語の語彙を削除した. 一般的な日本語の抽出には 3 つの言語資源(みんなの日本語の単語, 日本語の基本語彙 1000, 日本語能力試験 JLPT N1 単語集)を用いた.

5) 評価

(1) 放射線技術学用語集との比較

テキスト内に日本放射線技術学会から出版されている用語集(以下, JSRT 用語集)[5,6]に含まれる用語を抽出し, 本手法で同定できた用語の割合を算出した.

(2) 専門家による評価

JSRT 用語集に含まれなかった候補語について, 臨床経験 5 年以上の診療放射線技師 3 名が専門用語と認められる用語を選定した.

3. 結果

1143 語が候補語として抽出された. JSRT 用語集に含まれる用語がテキスト中に 792 語存在し, そのうち 309 語(39.0%)が候補語と一致した.

専門家による評価では, JSRT 用語集に含まれなかった 834 語のうち 1 人以上が選定した語は 820(98.4%), 2 名以上が 775 語(92.9%), 3 名全員が選定した語は 420 語(50.3%)であった.

Table.1 候補語の例

	用語
JSRT 用語集と一致	アレイコイル, イオン性造影剤, 安定同位体, 吸収線量
専門家 3 名が選定	傾斜磁場, 減弱, スライス厚, 照射線量, 断面積, 遺伝的影響
専門家 1〜2 名が選定	心臓, ガイドライン, 個人情報, 診療録, 頭部外傷, 死亡率
専門家が選定せず	距離, 温度, シフト, 数量, 異物, 告示, アインシュタイン, 拡大表示

4. 考察

本手法により JSRT 用語集に含まれていない 834 候補語のうち, 774 語(92.9%)を 2 名以上が選定したことから, 本手法は専門用語の抽出に有用であると考える. 3 名全員が選定した語は 420 語と約半数に減少した. この中の語の多くは, 解剖や疾患, 医療情報に関する語が多く, これらは放射線技術のみならず他の医療分野でも使用される. そのため, 専門家の中には放射線技術に特化した語ではないと判断されたことが原因であると考える. JSRT 用語集に掲載されている用語の抽出精度は 39%であった. 検出できなかった用語は, 候補語の調整において一般語として削除されてしまった可能性がある. 抽出精度の向上のためには, 今後さらなる検討を進めていく必要がある.

今後は, ひらがな, カタカナと漢字のみの用語を抽出対象としたが, アルファベットやその他の記号を考慮した手法についても引き続き検討していく. さらに, 提案手法は統計的情報のみを利用しているが, 今後は意味(概念)を考慮した用語を抽象化し, クラスタリングを行い、既存の用語集を拡大・改良することを目指す.

参 考 文 献

[1] 辻真太朗, 谷川原綾子, 福田晋久, 他. 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―. 日本放射線技術学会雑誌 .74(8) 757-76, 2018.

[2] 工藤拓. CaboCha /南瓜: Yet Another Japanese Dependency Structure Analyzer. [https://taku910.github.io/cabocha/ (cited 2019-Apr-18)]

[3] 関根聡. N グラム検索エンジン -Google 日本語 7 グラムを使って-, 言語処理学会第 14 回年次大会講演論文集(2008), A4-8

[4] Jabłoński A. From Complicated Into Simple: Declension in Japanese., *Silva Iaponicarum* 23/24/25/26, 2012.

[5] 日本放射線技術学会. 放射線技術学用語集. 京都:日本放射線技術学会, 1994.

[6] 日本放射線技術学会. 放射線技術学用語集・補遺編. 京都:日本放射線技術学会出版委員会, 2003.