



HOKKAIDO UNIVERSITY

Title	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
Author(s)	辻, 真太郎; 西本, 尚樹; 小笠原, 克彦
Citation	日本放射線技術学会雑誌, 64(7), 791-794
Issue Date	2008-07-20
Doc URL	https://hdl.handle.net/2115/39559
Type	journal article
File Information	ogasawara_JJRT64-7.pdf



原 著

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 —辻真太郎・西本尚樹¹⁾・小笠原克彦²⁾論文受付
2007年10月2日論文受理
2008年3月21日

Code No. 930

北海道大学病院放射線部

1)北海道大学大学院医学研究科医療情報学分野

2)北海道大学医学部保健学科医用情報科学分野

背 景

近年、病院情報システムの電子化が進み、電子化された医学文書(医学で生じるすべての文書を以下医学文書と定義する)が日々蓄積されている。放射線医学分野では、診療画像の読影レポート等が研究対象となっており、現在までに放射線読影レポートの形態素解析¹⁾や、形態素解析に基づいた用語間の意味関係の同定²⁾等の研究が行われている。形態素解析とは、「自然言語を適切な表現の境界で分割し、品詞情報を

付加させる処理」³⁾をいうが、形態素解析システムには多数の語が収録されており、文章中の語でこれらの語に該当しない品詞情報不明の「未知語」が課題となっている。医学文書では、記述形式が統一されていないことから、形態素解析の精度を低下させる要因となっている。形態素解析システムによって分析する場合、解析対象に含まれる多くの用語がシステム中の辞書に存在すれば未知語の語数は減少するが、無造作に辞書の登録語数を増加させていくことは時間的・人的コスト

Pilot Study of Domain-specific Terminology Adaptation for Morphological Analysis:
Research on Unknown Terms in National Examination Documents of
Radiological TechnologistsShintarou Tsuji, Naoki Nishimoto,¹⁾ and Katsuhiko Ogasawara²⁾

Department of Radiology, Hokkaido University Hospital

1)Department of Medical Informatics, Graduate School of Medicine, Hokkaido University

2)Department of Health Sciences, School of Medicine, Hokkaido University

Received Oct. 2, 2007; Revision accepted March 21, 2008; Code No. 930

Summary

Although large medical texts are stored in electronic format, they are seldom reused because of the difficulty of processing narrative texts by computer. Morphological analysis is a key technology for extracting medical terms correctly and automatically. This process parses a sentence into its smallest unit, the morpheme. Phrases consisting of two or more technical terms, however, cause morphological analysis software to fail in parsing the sentence and output unprocessed terms as "unknown words." The purpose of this study was to reduce the number of unknown words in medical narrative text processing. The results of parsing the text with additional dictionaries were compared with the analysis of the number of unknown words in the national examination for radiologists. The ratio of unknown words was reduced 1.0% to 0.36% by adding terminologies of radiological technology, MeSH, and ICD-10 labels. The terminology of radiological technology was the most effective resource, being reduced by 0.62%. This result clearly showed the necessity of additional dictionary selection and trends in unknown words. The potential for this investigation is to make available a large body of clinical information that would otherwise be inaccessible for applications other than manual health care review by personnel.

Key words: morphological analysis, radiology, unknown words, dictionary, medical documents

別刷資料請求先: 〒060-8648

札幌市北区北14条西5丁目

北海道大学病院 診療支援部(放射線部) 辻真太郎 宛

の点から現実的ではない。

今後、医学文書を対象とした研究を行うためには、分野が限定された文書の基礎的な解析を行い、辞書に追加する語の最適化を検討する必要がある。

そこでわれわれは、解析対象の文書を医学分野の枠組みに含まれる「放射線技術学」に限定するため、診療放射線技師試験問題(以下試験問題)を用いた。医学だけではなく、工学、物理学、化学、生物学など多くの領域の用語が含まれる試験問題を解析することで未知語の出現傾向を明らかにし、未知語の分析と低減化を検討した。

1. 方法

分析対象は、第54回(平成14年3月実施)および第55回(平成15年3月実施)試験問題のうち表・グラフ・画像に含まれる文字を除いたすべての文字(総文字数48,861字)とした。第56回(平成16年3月実施)以降は出題数、科目ごとの問題が定着していないため対象から除外した。形態素解析システムには「茶筌Ver2.3.3」[辞書IPADIC Ver.2.7.0(239,631語)⁴⁾を含む]を使用した。本研究では、IPADICに加え追加登録する用語群を追加辞書と定義し、以下の各追加辞書の組み合わせによる未知語を分類し、未知語率を比較した。

- (1)放射線技術学用語集[日本放射線技術学会(以下JSRT)、登録語4,670語]
- (2)MeSH日本語版(以下MeSH)(National Library of Medicine、登録語54,558語⁵⁾)
- (3)ICD-10対応電子カルテ用標準病名マスター(以下MEDIS)(財団法人医療情報システム開発センター、登録語31,153語⁶⁾)

未知語率(%) (以下 M_{Ratio})は以下で定義した。

$$M_{Ratio}(\%) = M_{Num} / N_{Total} \times 100$$

M_{Num} : 未知語数, N_{Total} : 検出された全語数

M_{Ratio} が低減するということは形態素解析システム中の辞書の用語と、分割された解析対象との語の一致が多くなることを意味する。

更に、領域ごとの未登録用語を把握するために、抽出された未知語を試験問題中の科目と対応させて集計し、各未知語の科目を特定した。なお、科目間で共通となる未知語はそれぞれ別の科目の未知語として扱うこととした。

2. 結果

放射線技術学用語集、MeSH、MEDISの3種類の辞書をすべて解析システムに追加した結果は、 $N_{Total}=32,557$, $M_{Num}=118$ より、 $M_{Ratio}=0.36\%$ となっ

た。辞書の追加がなかった場合、 $M_{Ratio}=1.0\%$ であるので、辞書の追加で M_{Ratio} は、追加前と比べて64%減少したこととなる(Table 1)。また、以下の用語が未知語と判別された。その一例を以下に挙げる。

リザーバーポート、腔、ホルツクネヒト、シンチレーション、インダクタンス、ラチチュード、ブラッグ、ボリウムレンダリング、ゾンネンカルプ、パナグラフィ、リニアックグラフィ、ベータ、クーリッジ、セルジンガー、ラクトパーオキシターゼ、径、ダイアグラム、等

科目別の M_{Num} をTable 2に示す。最も多かったのは電気電子工学の13語、次いで画像工学・エックス線撮影技術学11語、放射線計測学の8語であった。なお、記号、過剰な分割による漢字、半角文字はすべてその他に分類した。これはシステムが半角文字をすべて未知語として扱ってしまうこと、過剰な分割による漢字(すべて「腔」などの一文字の漢字)は単独で科目中の用語として意味を持っていないという理由からである。

3. 考察

3-1 追加辞書と未知語の関係について

各追加辞書の登録語数から未知語の低減効果を考えると、MEDISは約3万語の医学関連用語を含むにもかかわらず $M_{Ratio}=0.97\%$ であった(Table 1)。本研究で使用したMEDISは「病名」や「手術・処置名称」が中心であり、これらの用語が国家試験に含まれている用語と一致しなかったためと考えられる。西本²⁾らの研究結果ではこれらの辞書中に含まれる用語の重複の度合いが報告されている。単独で最も未知語を低減させた放射線技術学用語集と、最も低減に寄与しなかったMEDISとの収録用語の重複はMeSHと放射線技術学用語集等のほかの組み合わせと比較して少ないとの報告であった。更にCodan⁷⁾らの研究によると、大量の解析対象に対して、少数の用語を解析システムに追加することによって形態素解析結果が改善したという結果が報告されている。MEDISといった大規模な辞書よりも小規模である放射線技術学用語集が抽出結果を向上させたことはCodanらの研究結果と一致する。この結果から、追加する専門辞書を選択する際には、目的に合わせた追加辞書を選択する必要がある。

3-2 未知語の分類・分析

検出された未知語の周囲の文書構造から、未知語を以下の3種類に分類可能であると考えられる。

分類1: 検出された語が単独で概念としての意味を持つ

(例) エリアシング、ドライシルバ、レーゼ、ペルチェ、チョッパ、スラントホール

Table 1 Additional dictionary, the number of unknown words and the unknown word ratio.

Chasen (IPADIC)	Dictionaries			The size of dictionary (words)	The number of unknown words (M_{Num})	The unknown word ratio (M_{Ratio})
	JSRT	MeSH	MEDIS			
+	—	—	—	239,631	347	1.0%
+	+	—	—	244,301	204	0.62%
+	—	+	—	294,189	229	0.68%
+	—	—	+	270,784	329	0.97%
+	+	+	—	298,859	122	0.37%
+	+	—	+	275,454	184	0.56%
+	—	+	+	325,342	226	0.68%
+	+	+	+	330,012	118	0.36%

Table 2 Number of unknown words after three dictionaries were added.

The domain of word	The number of unknown words (%)
Electrical/Electronic engineering	13 (11.0)
Image engineering/Radiological imaging technology	11 (9.3)
Radiation detection and dosimetry	8 (6.8)
Nuclear medicine	7 (5.9)
Radiation physics	7 (5.9)
Radiological imaging equipment	6 (5.1)
Radiograph	6 (5.1)
Radiation therapy	5 (4.2)
Introduction to basis of medicine	3 (2.5)
Radiochemistry	3 (2.5)
Radiation control	2 (1.7)
Radiobiology	0 (0.0)
Other	47 (39.8)

分類2: 2語以上からなる複合語中の語が未知語となり概念が分割される

(例)... | 解析 | は | 読 | 影 | 者 | 間 | の...

肝癌 | の | 動 | 注 | リザーバーポート | 留置 | 術 |

分類3: 記号・半角英字

(例) Sonnenkalb, radioreceptorassay, repopulation,
√, Bq, °C, ±

分類1の場合, 「エリアシング」は辞書にはない新しい概念用語である。その他の用語では, 「方式」「法」「効果」「回路」「コリメータ」という語を付加することにより, 用語として成立する。分類2では, 複合名詞が過分割および形態素に分割される際の区切り方が問題となっていた。分類3では, 今回使用した解析システムでは, 半角の文字はすべて未知語として検出された。その他, 本研究で得られた未知語を分析すると, 同様の意味を表す用語であっても, 「スキヤッターリングホイール」と「スキヤッターリングフォイル」というよう

に外来語のカタカナ表記に表記ゆれ(ある事象に対して複数の表現が可能なこと, および表現が微妙に異なること)が存在していた。今後, 分割の際の区切り方や文字種別・表記ゆれを考慮した未知語の登録が必要であろう。

3-3 本研究の限界

本研究の限界として, 未知語の分類2でも明らかになった複合語の過剰な分割がある。西本²⁾らは, 一つの医学用語が形態素解析によって二つの概念に分割されることを指摘している。本研究でも本来検出されるべきである形態素が未知語ではなく二つまたはそれ以上の別の形態素に分割されている可能性があるであろう。例えば, 今回, 「液体シンチレーションカウンタ」という語は三つの追加辞書によって一般名詞と判別された。しかし, ほぼ同じ意味を持つ「液体シンチレーション装置」という語は「液体 | シンチレーション | 装

置」と分割され、「シンチレーション」という語が未知語と判別された。この解析結果に従って「シンチレーション」という語を辞書に追加すると解析結果全体として未知語は1語減るが、実際概念としての「シンチレーションカウンタ」を形態素として検出するのが難しくなる問題点を有している。このような不適切な分割を改善するために、現在、コーパスと呼ばれる、専門用語に対して「分類」や「クラス」等の情報を専門家により付与し、概念化する試みが行われている⁸⁾。コーパスは医学文書の解析には不可欠であるが、その作成には専門家の手作業による語への付属情報の付加が必要であり、多大な労力と時間を必要とする。更に、コーパスの精度を上げるためには、医療機関間および研究機関間による協力体制が不可欠であるだけでなく、読影レポート等では個人情報保護など倫理的問題があることから研究が進まないのが現状である。放射線技術学分野のコーパスは、放射線技術学領域の知識体系の構築に必要であり、JSRTが中心となった放射線技術学分野のコーパス構築が期待される。また、今回「電気電子工学」「画像工学・エックス線撮影技術学」「放射線計測学」等の分野別の分類は実際の試験問題と対応させた。そのため解析で得られた未知語は実際に臨床で使用されている用語と一致していない

可能性がある。これは放射線技術学特有の用語が存在し、未知語となっているためと考えられる。今後、電気電子工学などの専門領域の辞書の追加を検討したい。

4. 結 語

今回、本研究で以下の点が明らかになった。

- ・診療放射線技師試験問題を対象とした形態素解析を行った結果、3種類の辞書を追加した場合に最も未知語の数が減少した。未知語の内訳を国家試験の分類別に挙げると「電気電子工学」「画像工学・エックス線撮影技術学」「放射線計測学」の順で用語が不足していた。
- ・未知語は、新概念用語、過剰な分割や記号、表記ゆれによるものであった。形態素解析システムに辞書を追加する際は、文書の構成用語を理解する必要がある。

謝 辞

本研究を進めるにあたり、技術指導、貴重な助言を頂いた北海道大学大学院医学研究科社会医療管理学講座医療情報学分野の大学院生の皆様に大変感謝致します。

参考文献

- 1) 今井 健, 小野木雄三. 放射線読影レポートからの肯定・否定フレーム抽出の試み. 医療情報学 2003; 23 (Suppl): 484-487.
- 2) 西本尚樹, 寺江 聡, 蔭 国謙, 他. 複合語を構成する医学用語間の意味関係の同定における課題—CT画像診断レポートに出現する複合語を対象として. 医療情報学 2006; 25 (6): 413-420.
- 3) 吉村賢治. 自然言語処理の基礎 Information Science & Engineering-T4. サイエンス社, 東京, 2000: 4, 58.
- 4) Nara Institute of Science and Technology, Japan. URL: <http://chasen-legacy.sourceforge.jp/> (2008年6月24日確認)
- 5) U.S. National Library of Medicine, National Institutes of Health. URL: <http://www.nlm.nih.gov/mesh/meshhome.html> (2008年6月24日確認)
- 6) 医療情報システム開発センター (MEDIS-DC). URL: <http://www.medis.or.jp/> (2008年6月24日確認)
- 7) Coden AN, Pakhomov SV, Ando RK, et al. Domain-specific language models and lexicons for tagging. J Biomed Inform 2005; 38 (6): 422-430.
- 8) 荒木健治. 第1章 言葉をコンピュータで扱う方法, 第2章 コンピュータが言葉を覚える方法. 自然言語処理ことはじめ 言葉を覚え会話のできるコンピュータ. 森北出版, 東京, 2004: 13-20, 21.

図表の説明

Table 1 付加した辞書, 未知語数と未知語率

Table 2 3辞書を追加時の未知語数