

原 著

医学論文抄録を対象とした
略語表現自動抽出アルゴリズムの構築と評価若生理佳・西本尚樹¹⁾・上杉正人¹⁾・寺下貴美・小笠原克彦論文受付
2009年2月12日論文受理
2009年6月4日

Code No. 930

北海道大学医学部保健学科放射線技術科学専攻
1)北海道大学大学院医学研究科医療情報学分野

背 景

日常の医療現場では、非常に多くの医療用語が略語表現で使用されている。略語は、長い表現を短縮し、意思を効率よく伝えることができるため、多忙な医療現場では必須のものである。ところが、略語には以下のパターンが存在する。

- (1)一文字違いなどの似たような略語
- (2)同じ略し方でも異なった意味を複数持つ略語
- (3)施設依存に基づく勝手に作り出された造語のような略語

また、略語も時代とともに変化し、新しい略語も頻繁に誕生している。このような略語の使用はミスコミュニケーションを誘発し、医療従事者間における正

しい意思の疎通を阻害するため、診療放射線技師にとっては誤検査や誤照射などの医療事故の発生につながる可能性がある^{1,2)}。実際に、DIP (drip infusion pyelography: 点滴静注腎盂造影)とDIC (drip infusion cholecystochoangiography: 点滴静注胆嚢胆管造影)を混同し、造影剤の注入方法を間違えた造影剤注入ミスの事例や、照射録の略語による記載と読み間違いのために撮影部の左右を間違えて照射した誤照射の事例が報告されている¹⁾。また、Th7をC7と読み違え頸椎検査を行ってしまった誤検査の報告があり、撮影依頼伝票の略語による記載がその原因の一つとして挙げられている²⁾。

しかし、これまでの先行研究では、略語使用の間

Developing and Evaluating an Auto-retrieval Algorithm
for Abbreviations in Academic ArticlesMichika Wakoh, Naoki Nishimoto,¹⁾ Masahito Uesugi,¹⁾ Takayoshi Terashita, and
Katsuhiko Ogasawara

Department of Health Sciences, School of Medicine, Hokkaido University

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

Received February 12, 2009; Revision accepted June 4, 2009; Code No. 930

Summary

The purpose of this study was to develop an automated acquisition support for the semantics of abbreviations and evaluate its performance with respect to academic articles. Our goal was to support the maintenance of an institution-specific semantics inventory for abbreviations on a continuous basis. We retrieved articles from MEDLINE with the keyword "Liver [MeSH]," and 100 abstracts were randomly selected. Abbreviations and their full forms were retrieved using original Java software based on the following rules. (1) Searching the parentheses in the abstracts, the words inside the parentheses were retrieved as "INNER" and the words in front of the parentheses were retrieved as "OUTER." (2) Matching rules, such as whether the first characters of INNER and OUTER were the same. (3) If the words satisfied the conditions stated at (2), INNER was saved as the abbreviation and OUTER as the full form. Performance was manually evaluated by two graduate students and a radiologist. Of the 165 pairs of abbreviations and full forms that were obtained, 145(87.9%) constituted correct matches.

Key words: abbreviation, medical subject headings, MEDLINE, semantic inventory, rule-based programming

別刷資料請求先: 〒060-0812 北海道札幌市北区北12条西5丁目
北海道大学医学部保健学科 小笠原克彦 宛

題や標準的な略語表現辞書の構築方法に言及しているにすぎず、臨床現場における略語の知識共有を目指した研究は行われていないのが現状である³⁻⁵⁾。ステッドマン医学略語辞典など標準的な略語をまとめた略語辞書も出版されているが、実際の医療現場では辞書には載っていない略語も高頻度で使用されている^{1,2)}。

医療用語の標準化を進めることは、知識共有に必須である。しかしながら、既存の業務フローにのっている略語や専門用語を統一することは容易ではない。トップダウンで標準的な語彙を構築すると同時に、ボトムアップで既に使われている語をまとめあげ、標準的な語彙に接続する仕組みが必要である。

そこで、われわれは略語表現の自動抽出から知識共有に至るまでに以下の三段階を想定した。

(第一段階)既存の情報資源を使いながら個々の施設ごとに略語表現辞書を構築し、語義曖昧性の解消を行う

(第二段階)施設ごとに医療従事者間の知識共有を図る

(第三段階以降)施設間で略語表現辞書を比較し語彙の標準化、医療従事者間の知識共有を行う

研究の第一段階として、われわれは個々の病院に合った独自の略語表現辞書を自動構築することが、専門家同士の略語に関する知識共有を促進させ、略語が原因で発生する医療事故を防止することが必要であると考えた。また、自動構築を行うことにより、経時的に略語を収集し、時代の変化による略語の変遷という問題に対応可能である。

実際に、大阪大学医学部附属病院放射線部では、独自の略語集を作成している。試験的にRIS (radiology information system)に電子医学略語辞書を搭載し、この略語辞書が臨床現場でどの程度有用であるかを調査したところ、80%以上の回答者が有用であると答えている⁶⁾。

研究全体のゴールは、臨床現場において知識共有を促進することである。そこで、本研究では知識共有のなかでも略語表現における語義曖昧性を解消するために、略語表現辞書の自動構築手法を開発し評価を行った。

1. 方法および対象

1-1 解析対象

放射線読影レポートや退院時サマリーなど実際の臨床現場で使用されている文書には、略語とその基になる医学用語が同時に表記されない。われわれは、臨床医学文書に対応した略語表現自動抽出アルゴリズムを構築するには、核となる用語集を構築する

Table 1 Data characteristics: numbers of characters and words in the entire set of abstracts and the sample

	whole abstracts (5,000)	target abstracts (100)
Character counts	1251±377	1243±314
Word counts	212±66	210±51

ことが必要と考えた。同一の文章内で新出の略語にはその基になる医学用語が並列表記されるため、略語の解析対象には英語で記述された医学論文抄録とした。データベースはMEDLINE (PubMed)を使用して、“Liver”をMeSH統制語にて検索し、2007年11月14日から2008年11月14日までの1年間に掲載された医学論文抄録を取得した。得られた約5000件のうち100件をランダムサンプリングで選出し、実際に略語とそのフルスペルの抽出を行った。

取得した医学論文抄録約5000件とランダムサンプリングした100件の解析対象について、平均文字数と平均単語数をTable 1に示す。

1-2 アルゴリズム

抄録中に略語が最初に現れる際は、“full spell” (“abbreviation”)と表示されることが多い。ここで“abbreviation”とは略語表現を、“full spell”とはその略語の正式名称(以下、フルスペル)を意味する。この傾向を利用し、以下の手順にて略語の抽出を行った。

最初に、医学論文抄録中に登場するすべての括弧の中身を仮の略語(以下、INNER)として、括弧の直前にある語句を仮のフルスペル(以下、OUTER)として対で取り出した。ただし、フルスペルの語数は略語の文字数分、それ以下の単語数であると仮定し、OUTERとして括弧の直前から抜き出してくる語句の単語数は、INNERの文字数と同等とした(Fig. 1)。また、略語のフルスペルの最初の単語は、略語の頭文字から始まると仮定した。

INNERが略語であるかどうかの判定条件は、以下のとおりである。

- (1)文字数が2字以上10字以内である
- (2)頭文字がアルファベットの大文字である
- (3)ハイフン以外の記号を含まない

OUTERがそれに対応するINNERのフルスペルであるかの判定は次のように行った。INNERの頭文字と、OUTERの一語目の頭文字が一致するようにした。一致しない場合は、OUTERの一語目を削除し、OUTERの頭文字がINNERの頭文字と一致するまで同じ作業を繰り返した。一致するものがない場合は、略語とそのフルスペルの正しい組み合わせではない

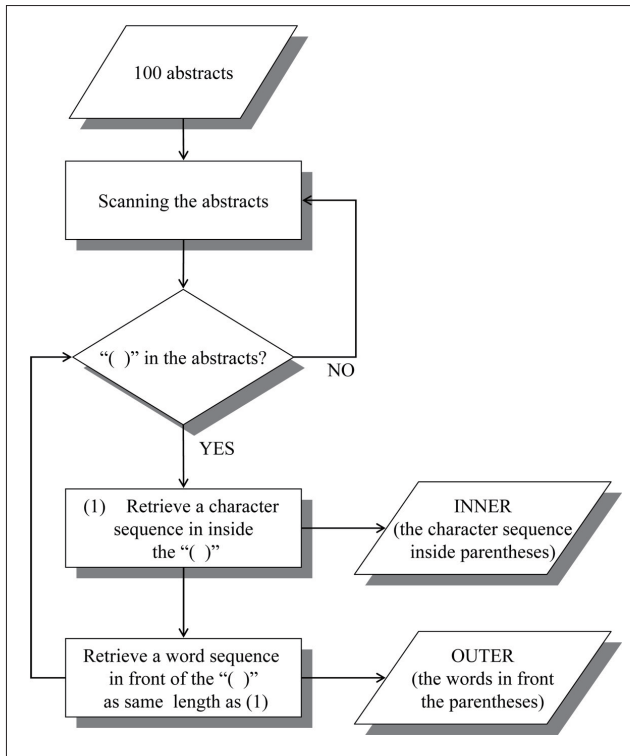


Fig. 1 Flow chart of the algorithm: INNER and OUTER word retrieval.

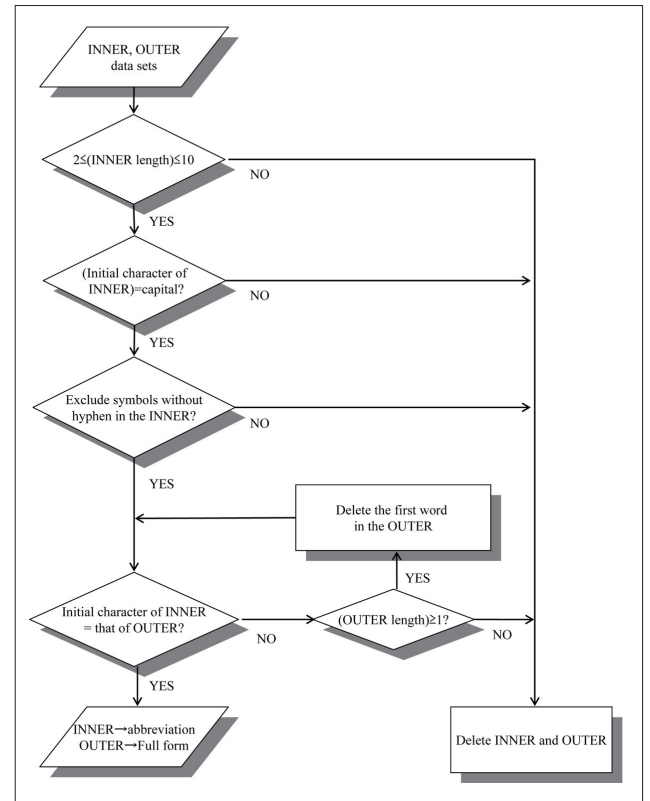


Fig. 2 Flow chart of the algorithm: conditional branch on INNER and OUTER.

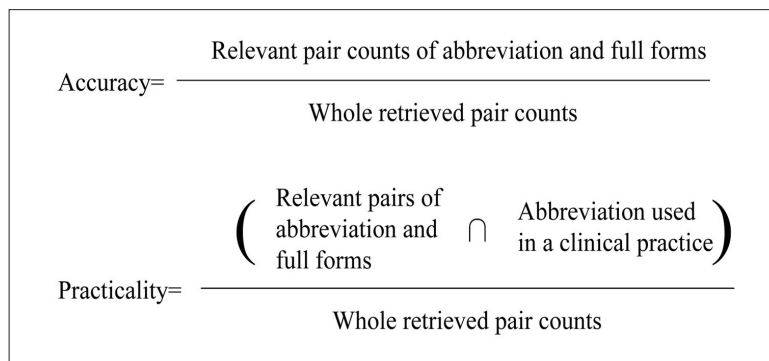


Fig. 3 Measurement of the algorithm.

と判断し、出力しなかった。

以上のルールをすべて満たした括弧の中身の文字列INNERを略語、括弧の直前にあった語句OUTERをその略語のフルスペルとして抽出した(Fig. 2)。また、抽出時、略語とそのフルスペル、抽出頻度、その略語とフルスペルを含む一文を同時に出力するようにした。

略語の抽出に用いる自作ソフトウェアの開発は、Java Standard Edition 6.0を使用し、統合開発環境はEclipse 3.2とした。

1-3 評価

評価は、アルゴリズムの精度と臨床医の略語使用経験による実用度について計算した。精度は、抽出

された略語とそのフルスペルの組み合わせのうち、略語とそのフルスペルの組み合わせとして、正当であると人間が判断した組み合わせの数が、全抽出数に占める割合とした。ここで、解析対象の論文を執筆した著者が独自に生成した略語も、略語として正当であれば正当な組み合わせとして評価した。実用度は、略語とそのフルスペルの組み合わせとして正当かつ実際に臨床現場で使用されていると判断される略語とそのフルスペルの組み合わせ数が、全抽出数に占める割合とした。精度は2名の大学院生、実用度は1名の放射線科医により、手作業により評価を行った。精度と実用度を、Fig. 3で式に示す。

Table 2 Accuracy and practicality of the algorithm

Accuracy	Practicality	Total pair count
87.6%	37.6%	
145 pairs	62 pairs	165 pairs

Table 3 Examples of abbreviations and full forms

ABBR ¹⁾	F ²⁾	Full form	Contexts	Rel ³⁾
ADRs	1	adverse drug reactions	PURPOSE: To determine the frequency of N-acetyltransferase 2 (NAT2) polymorphisms, the NAT2 acetylation profile and its relation to the incidence of gastrointestinal adverse drug reactions (ADRs), anti-tuberculosis (TB) drug-induced hepatotoxicity, and the clinical risk factors for hepatotoxicity in a population from Brazil.	○
AFS	1	Atomic fluorescence spectroscopy	Atomic fluorescence spectroscopy (AFS) was applied for mercury and selenium quantification.	○
ALT	2	alanine aminotransferase	Aspartate aminotransferase (AST), alanine aminotransferase (ALT), blood urea nitrogen (BUN) and creatinine levels and lactate dehydrogenase (LDH) activity were measured in the serum samples, while TNF-alpha, IL-beta, IL-6 and total antioxidant capacity (AOC) were assayed in plasma samples.	○
ALT	1	alanine transaminase	Trp-P-2 (3-amino-1-methyl-5H-pyrido [4,3-b] indole) ingestion for 42 d by C3H/HeJ mice caused elevation of serum alanine transaminase (ALT) activity and several signs of liver injury.	○
AMPK	1	an AMP-activated protein kinase	Sodium arsenite augments hepatic SHP mRNA levels in an AMP-activated protein kinase (AMPK) -dependent manner.	×
CRTH2	1	cells	RT-PCR revealed that three distinct PGD (2) responsive receptors, prostanoid DP receptor, chemoattractant receptor-homologous molecule expressed on Th2 cells (CRTH2) and thromboxane A (2) receptor (prostanoid TP receptor), were expressed in liver MFs.	×

1) Abbreviation, 2) Frequency, 3) Relevancy

2. 結果

全部で165組抽出され、そのうち、145組が略語とそのフルスペルとして正当な組み合わせであった(精度87.6%, Table 2)。また、抽出された略語とフルスペルの組み合わせのうち、略語とそのフルスペルとして正当かつ実際に臨床現場で使用されている略語は62組であった(実用度37.6%, Table 2)。実際に抽出された略語とそのフルスペルの一部を、Table 3に示す。

3. 考察

3-1 抽出可能な略語とストップワード

例えば、computed tomography (CT), left ventricular (LV)のように、フルスペルの単語数と略語の文字数が等しく、かつ、略語の頭文字とフルスペルの一語目の頭文字とが一致するものは、正しく抽出することが可能であった。同様に、hepatocellular carcinoma (HCC)のように、フルスペルの単語数が略語の文字数より少なく、かつ、フルスペルの一部に略語の頭文字から始まる略語と関連しない単語がないときは、正

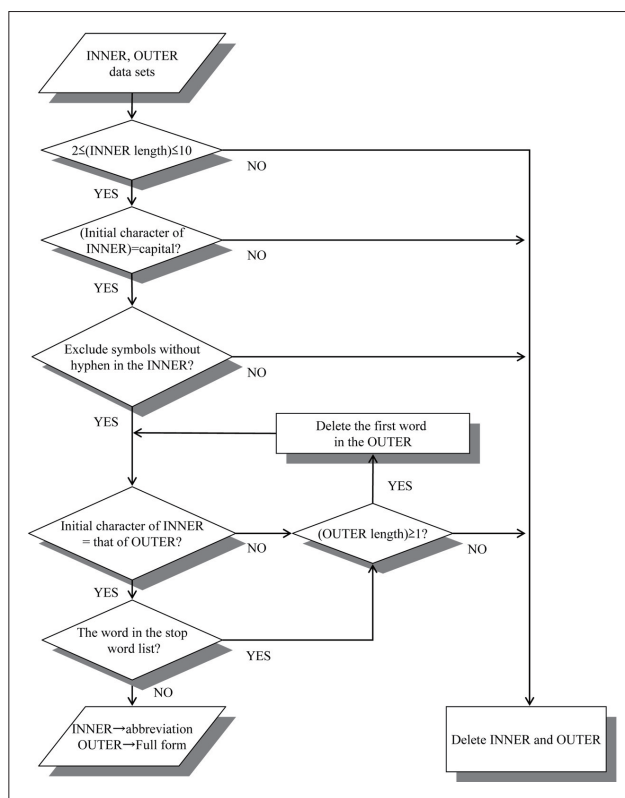


Fig. 4 Flow chart of the algorithm: stop word list injection.

しく抽出することが可能であった。一方, human hepatocellular carcinoma (HCC), an AMP-activated protein kinase (AMPK)のように, 略語に関連しない単語が存在する場合は, 正しい抽出ができなかった。

ここで, 誤って抽出されたan AMP-activated protein kinase (AMPK)に注目する。略語を構成する用語には, 通常, フルスベルがanのような冠詞から始まる略語はみられない。そこで, 除外されるべき語, いわゆるストップワードを設定することにより改善可能であると考えられる。日本語では助詞(例: て, に, を, は, が), 英語では冠詞, 前置詞, 接続詞など文中に高頻度で出現する語があるが, それらをインターネット検索における「検索キーワード」に設定すると検索結果が膨大になり検索効率が悪くなる。そのために, あらかじめ検索対象から除外させられている語をストップワードという⁷⁾。このストップワードを略語抽出用に設定し, フルスベルが特定の語(例: a, an, and, as, by, theなど)から始まることを防ぐように処理を加えることにより抽出精度が向上すると考えられる。

ストップワードを設定したときの具体的な処理内容は, 以下とした。略語の頭文字と, フルスベルの頭文字と一致した最初の単語がストップワードだった場合, そのストップワードだった単語は削除し, ほかに一致する単語はないか後続の単語を探索する。ストップワードではなかった場合は, そのまま処理を終

了し, フルスベルとして決定する。Fig. 4にそのアルゴリズムを示す。

仮に, ストップワードとしてanを設定した場合の処理を以下に示す。

- (1) 括弧内のAMPKは4字なので, 括弧の直前“an AMP-activated protein kinase (AMPK)”から4語を抜き出してくる。
- (2) anの頭文字とAを比較すると一致する。ここで, ストップワードであるanを削除する。
- (3) 新しく仮のフルスベルの一語目となったAMP-activatedの頭文字とAを比較すると, 一致する。ここで, AMP-activatedはストップワードではない。したがって, 処理を終了する。
- (4) AMP-activated protein kinaseがAMPKのフルスベルとして確定される。

AMP-activated protein kinaseは, 略語AMPKのフルスベルとしてふさわしく, ストップワードを設定することにより正しい抽出が可能となった。本手法において, 適切にストップワードを設定したと仮定して精度を測定し直したところ, 約90%に向上し, ストップワードを設定により略語に対しよりふさわしいフルスベルが抽出でき, 精度を上げることが可能となる。ただし, 略語の頭文字から始まる略語と関連しない単語として, ストップワード以外の単語が一致してしまう場合もある。例えば, HCCのフルスベルとして抽出

された、human hepatocellular carcinomaのhumanのように、その略語の一部を構成する単語として可能性がある場合は対処できない。この場合、辞書による語彙判定などが必要であるため、今後の課題である。

3-2 本研究で提案する手法の限界

本手法では、略語のフルスペルは略語の文字数と同一かそれ未満と仮定して抽出している。したがって、括弧の直前に出現する単語を探索し、フルスペルの単語数が略語の文字数よりも多い場合は抽出できない。改善策として、括弧の直前から抜き出してくる抽出語数を増やすことが挙げられる。しかし、抽出語数が増えると、それだけ略語に関連しない単語も多く抜き出され、精度が低下する可能性がある。これらを解決するためには、ストップワードとの併用により対応可能であると考えられる。更に、『OUTERにofが含まれる場合は、1語多く遡って抽出する』というルールも有効であろう。このようなルールを付加する場合、フルスペルには存在するがその頭文字が略語に存在しない語(例: of)について、研究を深める必要がある。

更には、blood flow rate (QB)のように略語の頭文字とフルスペルの頭文字が一致しない場合は、抽出できない⁸⁾。本手法では、略語の頭文字とフルスペルの頭文字との一致で略語とそのフルスペルとして抽出できるか判定しているため、括弧の直前から抜き出してくる抽出語数を増やしても、略語の頭文字とフルスペルの頭文字が一致しない略語には対応できない。このような例は稀であるが、表記のばらつきやソースの構築にも関わる問題であるため、今後検討が必要である。

3-3 実用度を低下させる要因

実際に臨床現場で使用されていると評価された略語は、精度(87.6%)に対して実用度は37.6%まで低下した。そのため、本研究で使用した手法と解析対象で得られた結果を、医療事故防止のためにそのまま臨床現場で役立てるにはまだ困難である。

実用度を低下させた原因として、以下の3点が挙げられる。

(1)略語とそのフルスペルの抽出対象に医学論文抄録を使用したことにより、抽出する略語の範囲が制限を

受けた

(2)医学論文抄録に出てくる略語と、実際の臨床現場で使われる略語が乖離している

(3)医学論文抄録を書く際に、著者が著者の意向により語句を略すことが多い

今後の改善点として、略語の抽出対象を医学論文抄録ではなく、実際の臨床現場を反映している臨床医学文書から略語とそのフルスペルの抽出を試みる必要がある。

3-4 臨床医学文書への適用

一般的に、実際の臨床医学文書では略語のみが記され、そのフルスペルまでは記述されないため、略語とそのフルスペルの抽出を別々の問題として扱う必要がある。臨床医学文書における略語の認識は、主にアルファベット大文字で表現される単語というルールで可能である。その略語に対するフルスペルの抽出であるが、フルスペルを臨床医学文書から得ることはほぼ不可能である。ここで、本手法であらかじめ略語辞書のある程度構築しておき、臨床医学文書から得られた略語を略語辞書で検索し、そのフルスペルを参照することは十分可能である。参照されるフルスペルが複数である場合、対応するフルスペルがない場合は、手作業でフルスペルを付加することも有効であると考えられる。

4. 結論

臨床現場における知識共有を促進するために、略語辞書の自動構築手法を開発し評価を行った。英語で記述された医学論文抄録から略語とそのフルスペルの自動抽出は、単純なルールに基づく機械的抽出と複数人による同時評価とを組み合わせた方式において、87.6%の精度で十分抽出可能であった。今後の課題として、適切なストップワードの設定による精度向上、ならびに、実際の臨床医学文書からの略語の自動抽出の必要性が明らかになった。

謝辞

本研究の実用度を検討するにあたり、快くご協力いただいた北海道大学医学部保健学科 西岡 健教授に深く感謝いたします。

参考文献

- 1) 熊谷孝三, 天内 廣, 太田原美郎, 他. 放射線業務における医療事故防止に関する学術調査 第二報 一般撮影, ポータブル撮影, 透視造影検査, 血管造影検査のリスク事例. 日放技学誌 2004; 60(6): 787-795.
- 2) 熊谷孝三, 天内 廣, 太田原美郎, 他. 放射線業務における医療事故防止に関する調査報告 第三報 CT検査, MRI検査, 核医学検査, 放射線治療のリスク事例. 日放技学誌 2004; 60(7): 927-938.
- 3) Ao H, Takagi T. ALICE: an algorithm to extract abbreviations from MEDLINE. J Am Med Inform Assoc 2005; 12(5): 576-586.
- 4) Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics 2006; 22(24): 3089-3095.
- 5) 岡崎直観, Ananiadou S, 辻井潤一. 共起情報に基づく医学文献からの略語抽出. 言語処理学会第12回年次大会 (NLP2006), 2006: 759-762.
- 6) 祐延良治, 笹垣三千宏, 佐藤和彦, 他. 放射線情報システム: 電子医学略語辞書の搭載とその評価. 日放技学誌 1999; 55(6): 588-594.
- 7) Savova GK, Coden AR, Sominsky IL, et al. Word sense disambiguation across two domains: biomedical literature and clinical notes. J Biomed Inform 2008; 41(6): 1088-1100.
- 8) Huang Z, Letteri JJ, Clark WR, et al. Ultrafiltration rate as a dose surrogate in pre-dilution hemofiltration. Int J Artif Organs 2007; 30(2): 124-132.

図表の説明

- Fig. 1 アルゴリズムのフローチャート: INNERとOUTERの抽出
Fig. 2 アルゴリズムのフローチャート: INNERとOUTERに関する条件分岐
Fig. 3 アルゴリズムの評価指標
Fig. 4 アルゴリズムのフローチャート: ストップワードを設定した場合

- Table 1 全取得抄録および解析対象における文字数と単語数
Table 2 抽出アルゴリズムの精度と実用度
Table 3 略語とそのフルスペルに関する抽出結果の一部