

一般口演

一般口演18

医療データ分析6（機械学習・テキストマイニング1）

2018年11月25日(日) 09:00 ～ 10:30 D会場 (4F 413+414)

[4-D-1-5] フリーテキスト形式の読影レポートにおける複合語パターンの特定

○張 洪健¹, 辻 真太郎^{2,3}, Wen Andrew³, 曹 瀛丹¹, Jiang Guoqian³, 小笠原 克彦² (1.Graduate school of health science, Hokkaido University, 2.Faculty of Health Sciences, Hokkaido University, 3.Department of Health Sciences Research, Mayo Clinic)

自由な形式で記述される放射線読影レポート(以下、レポート)には、疾患の有無やその解剖学的な位置等を記載するために複数の専門用語が組み合わせられて複合語として使用されるが、放射線科医による表現の違い等により一意に決めることは難しく、レポートから固有表現を抽出するのを妨げる要因の1つとなっている。また、先行研究では複合語の判断基準は読影医に委ねられており、どの程度複合語のパターンがレポートに存在するか明らかになっていない。そこで本研究では、英語の放射線読影レポートにおける複合語の実態を明らかにすることを目的として、医学分野の標準用語集における複合語の掲載状況を把握するとともに、用語集に収録されていない複合語の傾向を明らかにした。まず、データベースからランダムに抽出した英語の読影レポート400件を用いた。複合語の抽出には、医学分野の標準用語集である SNOMED-CTが実装されている cTAKES により、読影レポートに関する標準用語集である RadLexを統合して使用した。次にレポートに対して複合語のコーパスを作成し、ゴールドスタンダードとした。複合語の注釈は、名詞句内で最も右側にある語を被修飾語、その左側にある語を修飾語として考えられる全てのパターンの複合語をタグ付け、F値、適合率、再現率を求め、偽陰性となった複合語の構成語数と被修飾語を求めた。標準用語集のF値は、26.3% (適合率 45.5%, 再現率 18.5%) となった。撮像装置毎のF値は、CT(30.1%), X線(26.5%), PET(25.8%), MRI(22.3%)となった。また、偽陰性となった複合語は、多い順に3語(57.0%), 2語(28.8%), 4語(10.5%)で構成されていることが明らかになった。また、被修飾語で最も多用された語は順に uptake(592回), change(426回), disease(252回)であった。

フリーテキスト形式の読影レポートにおける複合語パターンの特定

張 洪健¹、辻 真太郎^{1,2}、Andrew Wen²、蔣 国謙²、曹 瀛丹¹、小笠原克彦¹

1 北海道大学大学院保健科学院、2 Mayo Clinic Department of Health Sciences Research

Grasping Patterns of Compound Terms for Mining Free-text Radiology Reports

Zhang Hongjian¹, Tsuji Shintaro^{1,2}, Andrew Wen², Jiang Guoqian², Cao Yingdan¹, Ogasawara Katsuhiko¹

1 Graduate School of Health Sciences Hokkaido University

2 Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, USA

Abstract

Extracting knowledge from free-text radiology reports is difficult because identifying compound terms has a variety of expressions described by radiologists. Thus, we tried to find the feature of the compound term in terminologies and improve the performance of the terminologies. We used cTAKES combined with RadLex to extract the compound terms from free-text radiology reports. We manually annotated all the patterns of compound terms and evaluated the performance of the terminologies using a recall, precision, and F-measure for training and validation steps. In the validation step, we combined the missing terms that didn't match with the dictionaries of cTAKES. We surveyed the feature of missing terms. While the most used stem word in the report of each modality was change(s)(CT and MRI), uptake(PET), and effusion(s)(X-ray). The F-measure of validation was increased by 31.4% compared with that of training (33.0% vs. 64.4%). The missing terms were mainly consisted of 2 to 4 words. The word "uptake" was the most extracted as the stem words. Although our results of validation step improved the performance but it still lower than that of similar research. This was caused by the method of annotation so that the performance could be improved in future.

Keywords: Radiology reports, compound terms, RadLex, terminology, free-text

1. 緒論

フリーテキスト形式で記述される放射線読影レポート(以下、レポート)には、疾患の有無やその解剖学的な位置等を記載するために複数の専門用語が組み合わされて複合語として使用されるが、放射線科医による表現の違い等により一意に決めることは難しく、レポートから固有表現を抽出するのを妨げる要因の1つとなっている¹。また、先行研究^{2,3}では複合語の判断基準は読影医に委ねられており、どの程度複合語のパターンがレポートに存在するか明らかになっていない。

2. 目的

本研究では、英語の放射線読影レポートにおける複合語の実態を明らかにすることを目的として、医学分野の標準用語集における複合語の掲載状況を把握するとともに、用語集に収録されていない複合語の傾向を明らかにした。

3. 方法

3.1 複合語の特定

レポート中で情報抽出の対象とされるのは主に名詞句であるため、本研究では最初にレポート中の名詞句を抽出した。次に、名詞句に対して stop word removal(SWR)を行った。最後に、名詞句内で構成される複合語のパターンを全て特定した。例えば、「There are additional small pulmonary nodes.」は名詞句として「There」と「additional small pulmonary nodes」を抽出し、次に、SWR を行い、最終的に「additional small

pulmonary nodes」を抽出する。この場合複合語は、「pulmonary nodes」「small pulmonary nodes」「additional small pulmonary nodes」となる。なお、本研究では、一番右に配置される被修飾語を語幹と定義した。

3.2 評価指標

評価指標は Recall、Precision と F-measure を使用した。Recall はレポートから抽出した複合語の中で、標準用語集に掲載されている複合語の比率であり、Precision は標準用語集を使って抽出した複合語の中で、手作業でも抽出した複合語の比率である。F-measure は Recall と Precision の重み付け平均値である。本研究は複合語が部分的に重複する場合を考慮した(下式)。

$$Recall = \frac{A + (0.5 \text{partial})}{A + C + (partial)} \quad Precision = \frac{A + (0.5 \text{partial})}{A + B + (partial)} \quad F_{measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

A: 標準用語集と手作業両方で抽出された複合語

B: 標準用語集のみで抽出された複合語

C: 手作業のみで抽出された複合語

3.3 対象

放射線読影レポートはオープンデータベースの MIMIC-III から抽出した。MIMIC-III は約 52 万件の英語のフリーテキストの放射線読影レポートを含むオープンデータバ

ースである。その中からランダムに CT, MRI, PET, X 線の各 125 件、総数 500 件の放射線読影レポートを抽出して、診断の情報を多く含む Findings, Impression, Interpretation を分析対象とした。その中で、教師データとして、総計 400 件(各画像検査 100 件)の放射線読影レポートを使用し、総計 100 件(各画像検査 25 件)を検証データとして使用した。

使用したソフトウェアは cTAKES ver4.0 である。実装した標準用語集には cTAKES のデフォルト用語集(SNOMED-CT、RxNORM)に加えて、放射線医学領域に特化した標準用語集である RadLex⁴とした。

3.4 複合語の特徴抽出と F-measure の算出

はじめに 400 件のレポートに対して cTAKES を用いて、標準用語集に掲載されている複合語を抽出した。同時にその 400 件の読影レポートに対して手作業で複合語を特定した。得られた二つの結果の重複を取り除いて、用語集に含まれない複合語を得た(以下 missing terms)。そして複合語の構成語数と頻度を明らかにして、F-measure を求めた。次に、得られた missing terms を用語集に追加して、残った 100 件のレポートの複合語を抽出した。100 件のレポートに対して手作業で複合語を特定し、F-measure を求めた。

4 結果

4.1 標準用語集に含まれない複合語の特徴

標準用語集に含まれない複合語の構成語数は、多い順に 2 語(56.9%), 3 語(28.7%), 4 語(10.5%)であった。また、語幹で最も多用された語は順に uptake(592 回)、change(s)(556 回)、lesion(s)(398 回)、node(s)(379 回)であった(表 1)。画像撮像装置毎に見てみると、CT で collection(s)、MRI で change(s)、PET で uptake、X 線で effusion(s)となった(表 2)。

表 1 出現頻度の高い missing terms の語幹(上位 10 位)

単語	回数(%)	単語	回数(%)
Uptake	592(4.2%)	Nodule(s)	235(1.6%)
Change(s)	556(3.9%)	Mass	201(1.4%)
Lesion(s)	398(2.8%)	Evidence	197(1.4%)
Node(s)	379(2.7%)	Effusion	187(1.3%)
Disease	252(1.8%)	Study	180(1.2%)

表 2 各画像撮影装置のレポートに最も多く含まれていた複合語の語幹

CT	MRI	PET	X 線
change(s)	change(s)	uptake	effusion(s)
collection(s)	enhancement	node(s)	opacity(ies)
lesion(s)	lesion(s)	lesion(s)	tube(s)
effusion(s)	artery(ies)	avidity	pneumoth-rax
evidence	narrowing	disease	line(s)

4.2 評価

教師データの F 値は、33.0% (precision 73.2%, recall 21.2%)となった。画像撮像装置毎の F 値は、CT(36.6%)、X 線(32.3%)、PET(31.9%)、MRI(30.7%)となった。検証データでは、すなわち missing terms を追加した標準用語集の F 値は、64.4%(precision 75.6%, recall 56.1%) となった。画像撮像装置毎の F 値は、CT(62.8%)、MRI(65.0%)、PET(66.2%)、

X 線(64.0%)となった。

表 3 検証データにおける画像撮像装置毎の増加量

	CT	MRI	PET	X 線	全体
Recall	29.6	35.6	37.5	36.8	35.1
Precision	2.9	5.2	4.9	2.5	2.5
F 値	25.8	34.0	34.2	31.9	31.4

5 考察

本研究は放射線読影レポートから標準用語集に基づいた複合語と手作業で特定した複合語の比較を通じて、用語集に含まれない複合語の特徴を明らかにした。用語集に収載されていない複合語では、PET の語数が多く、デフォルト用語集と RadLex が PET に関する複合語をあまり収載していないことが推測される。missing terms を用語集に追加すると、Recall, Precision および F-measure が上昇するが、類似研究の成績と比較すると低い⁵。この原因としては、手作業の複合語の特定方法が半自動のため、複合語として意味をなさないパターンを含み、Recall を低下させている可能性が考えられる。また、先行研究の複合語は、放射線科医が決定しており、我々の特定した複合語より総数は少ないことが予想されるため、同様に Recall の上昇の余地があると思われる。今後は、本研究で構築した辞書を実際の読影医師により得られる複合語と比較・検証を行うことで、フリーテキスト形式の読影レポートからの知識抽出が一層改善されると考える。

6 結論

本研究では、フリーテキストの放射線読影レポートを対象として、標準用語集に含まれない複合語の特徴として構成語数と語幹を明らかにした。また、画像撮像毎の F-measure の傾向を明らかにすることで、標準用語集に収載する語の範囲を検討する基礎資料を提供した。本研究で構築した複合語の辞書を用い、今後実際の放射線科医から得られる複合語と比較することで、類似研究との比較が可能となると考えられる。

参考文献

- 1) <https://www.ezdi.com/resources/crf-based-clinical-named-entity-recognition-using-clinical-nlp-features/>(2018/9/5 確認)
- 2) Dirk Marwede, Indexing Thoracic CT Reports Using a Preliminary Version of a Standardized Radiological Lexicon (RadLex), Journal of Digital Imaging, 363-370, 2008
- 3) Ryan W.Woods, John Eng, Evaluating the Completeness of RadLex in the Chest Radiography Domain, Acad Radiol, 1329-1333, 2013
- 4) https://www.rsna.org/RadLex_Playbook.aspx(2018/9/5 確認)
- 5) Yang Huang, Henry J.Lowe, Dan Klein, Russell J.Cucina, Improve Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon, Application of Information Technology, 275-285, 2005