一般口演 | 知識工学

一般口演10
知識工学

2019年11月23日(土) 09:00 〜 11:00 C会場 (国際会議場 2階国際会議室)

[3-C-1-07] 放射線医学領域の標準用語集を拡張する機械学習のパラメータの検討

○張 洪健[1]、辻 真太朗[2]、Andrew Wen[2]、蒋 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 College of Medicine, Department of Health Sciences Research）

キーワード：Radiology Report, Radlex, word2vec, NER, technical term

【背景・目的】放射線読影レポートにおける標準化用語集を利用した辞書ベースの固有表現抽出は、抽出した複合語間の関連関係を表示できることなどの利点があるが、ルールベースや機械学習を用いた手法と比較すると、その精度は高いとは言えない。そこで、放射線領域で代表的な標準用語集・ Ontologyである RadLexを使用した辞書を拡張するためには、機械学習である Word2Vecを用いて、辞書拡張に対して最適な類似語を生成する必要がある。本研究では最適な類似語の生成に必要な Word2Vecの最適なアルゴリズムとパラメータを検討する。

【方法】本研究はオープンデータベースである MIMIC-III内の放射線読影レポートの Findingsと Impressionの163,201件に対して、 Word2Vecを適用し、辞書を拡張する語を抽出した。更に、 Word2Vecが辞書拡張に対して最適な類似語を得られる様なアルゴリズムとパラメータを検討した。抽出した語における頻度が一番高い5つの被修飾語を入力し、出力の中で cosine距離が0.5以上の被修飾語を付く複合語の数が多くなると、アルゴリズムの効果がよいと定義した。検証したアルゴリズムは以下の3種類、 CBOW(以下 NS-CB)、 Skip-Gram(以下 NS-SG)、 Skip-Gram(以下 HS-SG)とした。次に、効果が最もよいアルゴリズムに対して、 epoch数が3から9まで設定して、最適な epoch数を検討した。

【結果】出力の中で類似語の数は、 NS-CBが全て0、 NS-SGが最高10、最低0、 HS-SGが最高13、最低1になった。アルゴリズムは HS-SGの場合で、 epoch数が7になる時に、 cosine距離の総和が極大値であった。そのため、アルゴリズムが HS-SGで、 epoch数が7となる場合、最適な類似語が得られると考えられる。

第 38 回日本医療情報学連合大会 詳細抄録・原稿テンプレート

放射線医学領域の標準用語集を拡張する機械学習のパラメータの検討

張 洪健[1]、辻 真太朗[1,2]、Andrew Wen[2]、 蒋 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

Exploring the parameter of machine learning for expanding the terminology of radiology medical

Zhang Hongjian[1] , Tsuji Shintaro[1,2], Andrew Wen[2], Jiang Guoqian[2], Cao Yingdan[1], Ogasawara Katsuhiko[1]
1 Graduate School of Health Sciences Hokkaido University
2 Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, USA

Abstract

Dictionary-based named entity recognition (NER) with standardized terminology in radiology reports has the advantage of expressing the association relationships between extracted compounds. An accuracy with NER is known lower than machine-learning approach such as conditional random fields (CRF). However, it has a benefit to performing NER based on standard vocabularies like ontology because relationships between technical terms are available. Therefore, we attempt to expand the RadLex as a terminology, which is a representative technical terms having agreement with radiologists. While grasping the trend of the words embedding in radiology reports, technical terms not identified by RadLex were added to the customized dictionaries in NER tool, and we cleared the accuracies of these terms. 163,201 items of diagnostic findings and impressions in MIMIC-III were used to extract words for expanding dictionaries using Word2Vec. The parameters of Word2Vec for enhancing ability of extracting the most similar words are discussed. The synonym with 7 epochs is the most optimized in our approach, which is used the hierarchical softmax based skip-gram algorithm. We proposed insights for improving dictionary-based NER using Word2Vec approaches such as building models, embedding of modifiers of compound words, and appending compound words to the dictionary associated with the order of output cosine values.

Keywords: Radiology reports, RadLex, Word2Vec, named-entity recognition (NER), MIMIC-III

1. 緒論

放射線読影レポートにおける標準化用語集を利用した辞書ベースの NER(Named Entity Recognition) は、抽出した複合語間の関連関係を表示できることなどの利点があるが、ルールベースや機械学習を用いた手法と比較すると、その精度は高いとは言えない。また、放射線領域で代表的な標準用語集・Ontology である RadLex を使用して辞書の語彙を拡張するためには、類似語を生成する必要がある。そこで本研究では，類似語の生成の最適化条件を探るために機械学習の手法である Word2Vec を用いて，放射線読影レポートにおける最適な用語生成のアルゴリズムとパラメータを検討することを目的とした。

2. 方法

2.1 複合語のパターンの定義

放射線読影レポートの中で「病名」や「疾患」、「解剖学的位置」などが記述されるのは、主に名詞句であり，句内の複合語の存在パターンには特徴がある。例えば、名詞句「the superior segment of the right lower lobe」を例に考えと，前半の the superior segment と、後半の the right lower lobe は複合語であるが，the superior segment of the right lower lobe も複合語と捉えることもできる. 放射線読影レポートを用いた先行研究では，長い名詞句を対象とした分析が行われてきたが，本研究では、修飾語・被修飾語に注目するために名詞句の最小単位である the superior segment と the right lower lobe の中の構造に注目して複合語のパターンを定義した。例えば、the right lower lobe は、right lower lobe と lower lobe の複合語が考えられる。この時、一番右側に配置されている語を本研究では「語幹」と定義した。また、語幹の左側に並ぶ語を「修飾語」と定義した。

2.2 分析手順

初めに、正規表現を用いて MIMIC-III 内の放射線読影レポートから Findings と Impression の部分を 163,201 件抽出した。次に、抽出したデータに対して、2.3 で示す前処理を行った。前処理したデータをトレーニングデータとし、Word2Vec を使用して、Negative Sampling に基づいた CBOW 法、Negative Sampling に基づいた Skip-Gram 法、Hierarchical Softmax に基づいた Skip-Gram 方法でトレーニングし、モデルをそれぞれ構築した(以下、NS-CB モデル、NS-SG モデル、HS-SG モデル)。Word2Vec のパラメータとして Window サイズを 5、word vector の次元数を 100、初期学習率を 0.025、単語の最低出現回数を 1、epoch を 5 に設定した。そして、それぞれのモデルに対して、頻度が高い語幹を入力し、類似する語を出力するとともに、各モデルの性能を比較した。次に、性能が一番良いモデルに対して、epoch 数を調整し、効果が最も良い epoch 数を明らかにした。

2.3 前処理

はじめに、前処理としてレポート内に含まれる段落のタイトル(Findings と Impression)や数値を含む表現(1.5×2.3cm など)を、複合語の生成にとってその精度を低下させるノイズ情報となる表現として特定・除外した。さらに、True Positive と False Negative となった複合語中のスペースをアンダーバー(_)に変換し、複合語を一連の単語として識別させる処理を行った。例えば、”lung lobe”と言う複合語を”lung_lobe”に変換して、それを一つの単語として処理した。最後に、対象の文書を小文字に変換して、プログラムを開発し Stop Word と句読点を除く処理を行った。

2.4 モデル効果の判断

本研究はアルゴリズムの最適性を判断する際に、語幹を Word2Vec の入力とし、出力結果内の cosine 値が 0.5 以上の結果の数をモデル効果に判断した。その数が多くなる程、モデルの精度が改善されたと判定した。そして、最適な epoch 数を判断する際には、各入力された語幹に対する出力結果の最大値の cosine 値の総和を比較し、高い方は精度が高いと判断した。

3 結果

各モデルの出力結果を表 1 に示す。表の数字は各語幹を入力とする際に、出力の上位 30 個の中で、入力語を語幹とする語の数であり、括弧内の数字は入力を含んでいる語の数である。表の結果により、HS-SG モデルの精度が高いことが明らかになった。

表 1　各モデルの比較

	NS-CB	NS-SG	HS-SG
Effusion	0	10(11)	13(17)
Change	0	0(1)	10(12)
Node	0	6(8)	11(14)
Lesion	0	1(2)	11(14)
uptake	0	3	1

HS-SG モデルに対して、epoch が 3~9 となる場合の各出力結果を以下に示す(図 1、図 2)。図 2 は各 epoch 数に対して、各入力(uptake を除外した)の上位 1 位の cosine 距離の総和を表す。

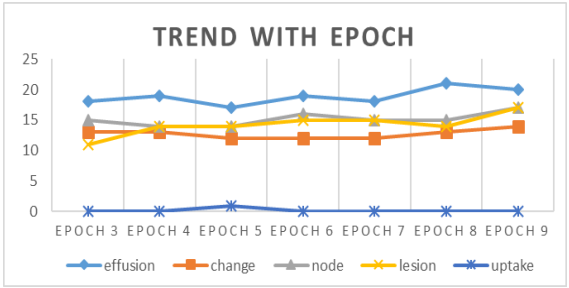

図 1　出力語数の変化

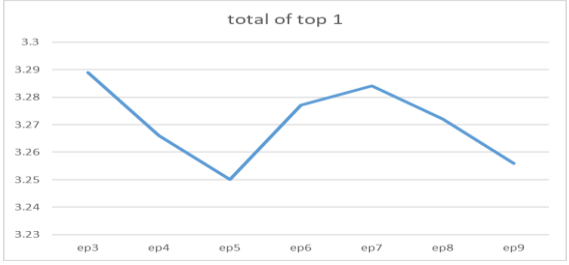

図 2　cosine 距離の総和の変化

4 考察

アルゴリズムは Hierarchical Softmax に基づいた Skip-Gram 方法にする場合に精度が最も高く、続いて HS-SG モデルに対してパラメータの影響を検討した。今回の研究は epoch 数による影響も検討した。他のパラメータは固定して epoch 数を 3 から 9 まで変動された(図 1～2)。図 1 は epoch 数が 3～9 の場合で、各語幹を入力とした時の出力の中で、それぞれを含んでいる語の数を表した。図 2 は epoch 数が 3～9 の場合で、各語幹(uptake を除外した)の一番目の出力結果の cosine の総和の変化を表した。語幹を含んでいる語数は epoch 数の増加によって増加したが、全体の cosine 値は epoch 数が 7 を超えた以後に減少したため、今回のパラメータにする場合に、epoch 数 7 が最適だと判定した。

本研究は辞書ベース NER の抽出精度を向上させるために、辞書に追加する必要性が高い語を提示したが，以下の 2 つの課題が明らかとなった。1 つは、Word2Vec のアルゴリズムとパラメータの 1 つである epoch 数の選択を検討したが、他のパラメータ、例えば次元数の影響がどの程度があるかは、明らかになっていない。2 つ目は、今回使われたトレーニングデータにとって、ノイズ情報、Stop Word と句読点の除外以外の前処理によって精度が改善する余地が残されていると考えられる。今後は、辞書ベースの NER 抽出精度を向上させるために、他のパラメータの最適化を検討する。

5 結論

本研究は，辞書ベース NER の抽出精度の向上を図るため、機械学習の 1 つである Word2Vec を用いて放射線読影レポート内の出現頻度が高い語幹を入力とし、語幹との類似性が高い複合語を得た。そして、類似性が高い複合語を得る効果をよくするために、Word2Vec の最適なパラメータを検討した。アルゴリズムは Hierarchical Softmax に基づいた Skip-Gram 方法を使用し、epoch 数を 7 にする場合に，類似する複合語を得る精度が高かった。本研究では、類似性が高い複合語を得る効果が最適化されるアルゴリズムと epoch 数を示したが、word vector の次元数などパラメータに関しては今後検討の余地があると考えられる。

参考文献

1) Dirk Marwede, Indexing Thoracic CT Reports Using a Preliminary Version of a Standardized Radiological Lexicon (RadLex), Journal of Digital Imaging, 363-370, 2008

2) Ryan W.Woods, John Eng, Evaluating the Completeness of RadLex in the Chest Radiography Domain, Acad Radiol, 1329-1333, 2013

3) T. Mikolov, J. Kopecky, L. Burget, O.Glembek, J.´Cernock˘y. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems.2013.

4) Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781(2013).