GunNLP at the NTCIR-16 Real-MedNLP Task:

Collaborative filtering-based similar case identification method via structured data “case matrix”

Rei Noguchi, Gunma University Hospital

NATIONAL UNIVERSITY CORPORATION Gunma University

Introduction

- The clinical text data such as radiology reports are highly expected to be utilized directly for diagnosis support or similar case search using natural language processing and machine learning model.
- Especially in medical fields, the explainability of a machine learning model is extremely important.
- In this study, I propose an explainable case identification model for radiology reports by structuring the reports into a “case matrix” and by applying a collaborative filtering algorithm.

Method

- I tried to apply **user-based collaborative filtering**, which is mainly used for personalized recommendations, to case identification.
- For the application, given training data of free text description were processed into **structured data “Case Matrix”**, in which each record was unique for each case, and the presence or absence of each symptom was stored in separate columns.
- **Normalization of spelling variants** using Manbyo-Dictionary, **detection of negation** and its simple dependency relationship, and **exclusion of stop words** were also conducted.

Scheme of developing “Case Matrix”

Radiology reports by different radiologists

MedTxt-RR-JA-training.xml

```
<articles>
<article id="0" title="03">
<a>左舌区</a>に<a>気管支区域に沿った</a><d
certainty="positive">consolidation</d>があります。<a>
内部</a>に<d certainty="positive">気管支拡張像</d>
 …
</article>
<article id="1" title="09">
<a>右肺下葉気管支区域に沿う</a><d
certainty="positive">浸潤影</d>があり、<d
certainty="positive">肺癌</d>と考えます。
<a>葉間胸膜と広く接し</a>ており、<d
certainty="positive">胸膜浸潤</d>ありと考えます。
 …
```

Data Processing

- Structuring data
- Dividing into words
- Normalizing spelling variants
- Detecting negation
- Excluding stop words

Developing “Case Matrix” via BoW (Bag of Words)

Art. ID	Case	腫瘤	肺癌	胸部異常陰影	肝	腫瘤(⊘)	浸潤	内部	癌	…
0	03	0	1	0	0	0	0	1	0	
1	09	2	2	1	1	0	1	0	0	
2	12	2	1	0	1	2	0	0	1	
3	03	1	0	2	0	0	1	1	1	
:	:	:								

Results

- For training data, a matching rate between the inferred similar case numbers and true ones was **60% (43 of 72)**.
- Dimensionality reduction by sparse modeling decreased the variables from 782 to 55 and improved the matching rate to **71% (51 of 72)**.
- Moreover, this method was able to **provide the basis for the inference by the model** based on the matched words.
- In contrast, for test data, the evaluation result notified by the organizer was 0.3569 in terms of normalized mutual information.

Scheme of collaborative filtering-based similar case identification

Case Matrix Feature words in case reports

Art. ID	Case	腫瘤	肺癌	胸部異常陰影	肝	腫瘤(⊘)	浸潤	内部	癌	浸潤影	胸膜	…
0	03	0	1	0	0	0	0	1	0	0	0	
1	09	2	2	1	1	0	1	0	0	1	0	
2	12	2	1	0	1	2	0	0	1	0	0	
3	03	1	0	2	0	0	1	1	1	1	2	
:	:	:										

Calculating correlation coef.

Case Vector (a case to be predicted)

80	?	1	0	1	0	0	1	0	1	1	0

Extracting similar cases base

Rank	Art.ID	Case	Num. of matched words	Corr. coef.
1	3	03	4	0.46
2	0	03	3	0.35
3	2	12	2	0.25
:	:		:	:

Matched feature words
(the basis for the inference by the model)

腫瘤　胸部異常陰影　浸潤　浸潤影

Conclusion

- The case matrix and collaborative filtering-based method worked for case identification **with explainability** in radiology reports.
- However, the accuracy for test data was much lower than that for training data. It might be because of the “cold start problem”, the high dimensionality, and the sparseness of the case matrix.
- I will consider the combination with model-based methods such as non-negative matrix factorization to improve accuracy and enhance functions such as the prediction of symptoms and complications.