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
rnoguchi@gunma-u.ac.jp

ABSTRACT

Clinical text data are highly expected to be utilized directly for medical examination or diagnosis to support doctors' practices. In this study, I propose a framework of similar case identification for radiology reports by structuring the reports into a "case matrix" and by applying a collaborative filtering algorithm.

KEYWORDS

Case identification, Case matrix, Collaborative filtering

TEAM NAME

GunNLP

SUBTASKS

Subtask3-RR-JA (CI)

1 INTRODUCTION

A lot of text data are recorded in electronic medical records such as radiology reports and have important and consistent information about patients and their clinical conditions. The clinical text data are highly expected to be utilized directly for medical examination or diagnosis to support doctors' practices or reduce medical disparities. In particular, oversight or misdiagnosis of important findings has frequently occurred in radiology reports recently, and therefore the text data should be utilized for medical support to reduce the incidents.

I have previously developed a method for extracting structured data "case matrix" from clinical text descriptions. In this study, I propose a framework of similar case identification for radiology reports [1] by structuring the reports into a case matrix and by applying a collaborative filtering algorithm.

2 RELATED WORK

I previously developed a method for extracting a case matrix, in which each record was unique for each case and the presence or absence of each symptom was stored in separate columns, from discharge summaries [2]. In addition, I presented a framework of collaborative filtering-based similar case matching via the case matrix for discharge summaries [3].

3 METHODS

3.1 Extraction of case matrix

To develop a case matrix from the radiology reports, I first processed the reports into structured data unique for each article and segmented the individual articles into words (Figure 1). After data cleansing and elaboration such as normalizing spelling variants, detecting negation relationships, and excluding stop words, I develop the case matrix by summing the appearance frequencies of the entities (mainly symptom expressions) for each article (Figure 2).

The details of word segmentation and data cleansing are described in the following sections.

3.2 Word segmentation by IRIS NLP technology

I used IRIS NLP technology [4] (also known as "iknowpy" library in Python) for word segmentation instead of morphological analysis. This technology can extract a word or group of words, called "entity", based on the grammatical structure without dictionaries (Figure 3). Since clinical texts include many compound words and technical words, dictionary-based morphological analysis has a limitation in the accurate word segmentation. In contrast, IRIS NLP is expected to segment words into appropriate units despite neologism due to the grammatical structure-based approach.

3.3 Data cleansing and elaboration

Clinical texts include many spelling variants (e.g., "咳", "せき", "咳嗽"), and these variants have a large influence on the variety of columns in the case matrix, leading to a decrease in the accuracy of case identification. I normalized the variants using "Manbyo-Dictionary" [5], a large-scale disease dictionary to associate disease names actually used in clinical texts with standard disease names. In addition, there are some negative findings (e.g., "咳は認められない") in the reports, and they should be detected and reflected on the case matrix. IRIS NLP technology can detect just the presence of negation, and therefore, I additionally developed an original simple algorithm to identify negation dependency relationship as much as possible.

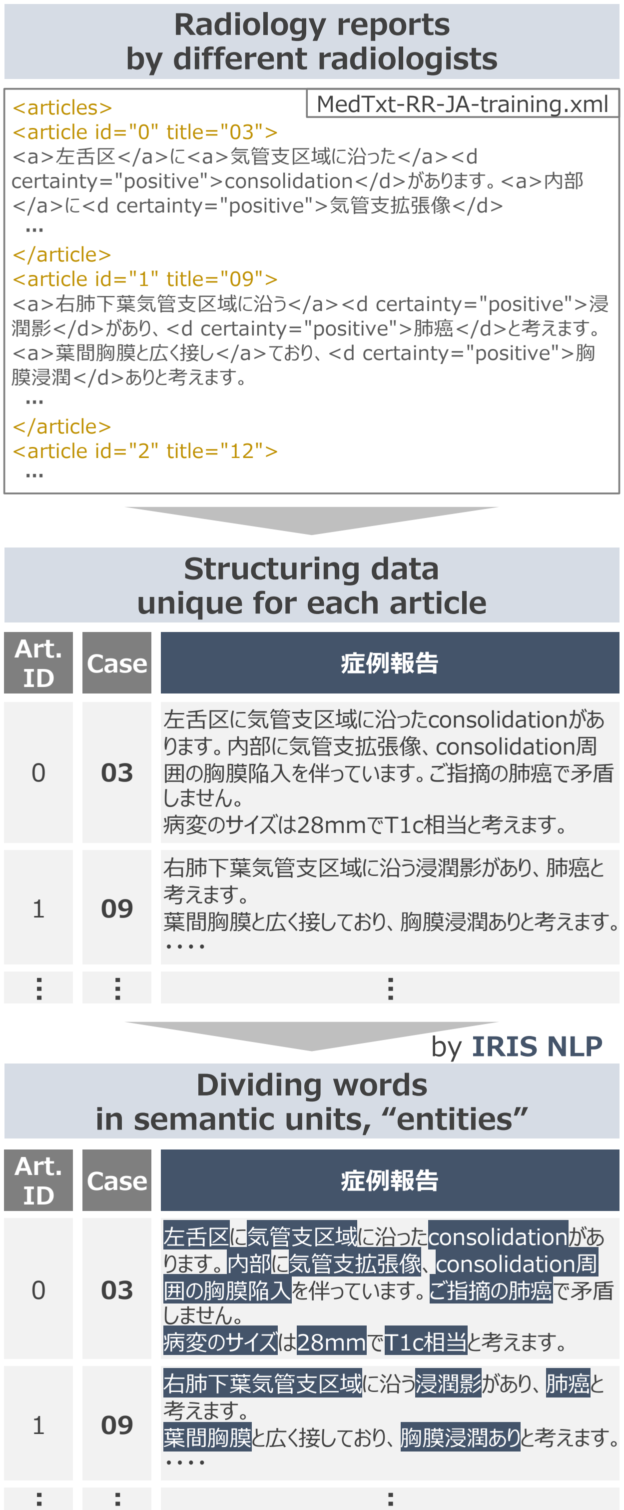

Figure 1: First step of data processing for developing “case matrix”

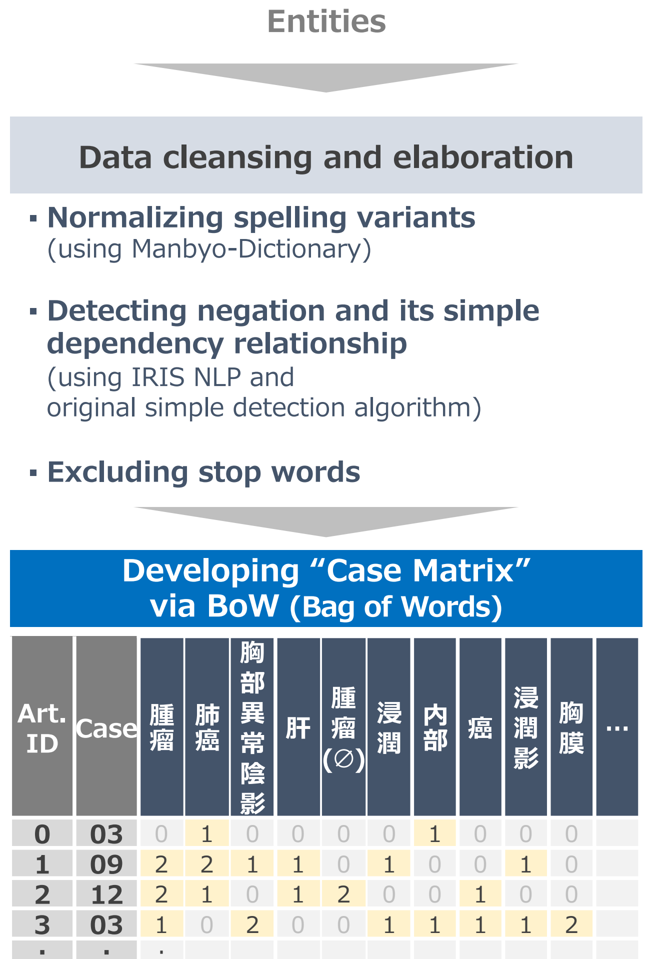

Figure 2: Final step of data processing for developing “Case Matrix”

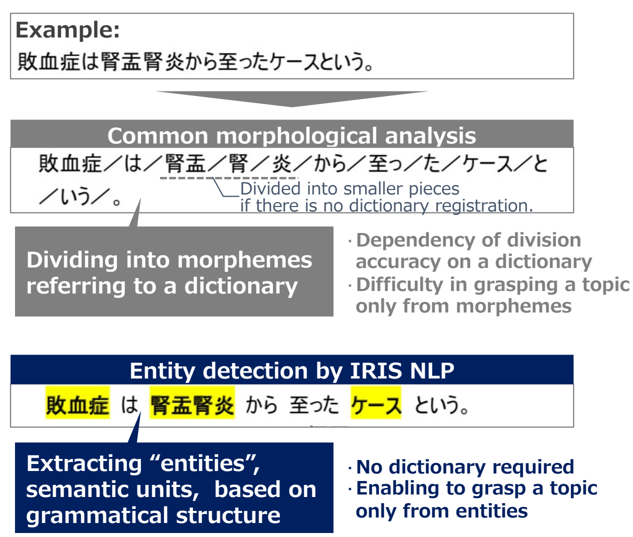

Figure 3: Overview of “entity” detection by IRIS NLP technology

3.4 Case identification by collaborative filtering

"Case matrix" is a useful data format for training data in machine learning, enabling us to utilize text data like numerical data for data analysis or machine learning. For case identification, I applied user-based collaborative filtering, which is mainly used in marketing fields for personalized recommendation systems in e-commerce sites [6-7], to the case matrix and also developed an extraction algorithm of similar cases for an unknown input case (Figure 4).

For similarity indexes, correlation coefficient, cosine similarity, and Hamming distance were considered.

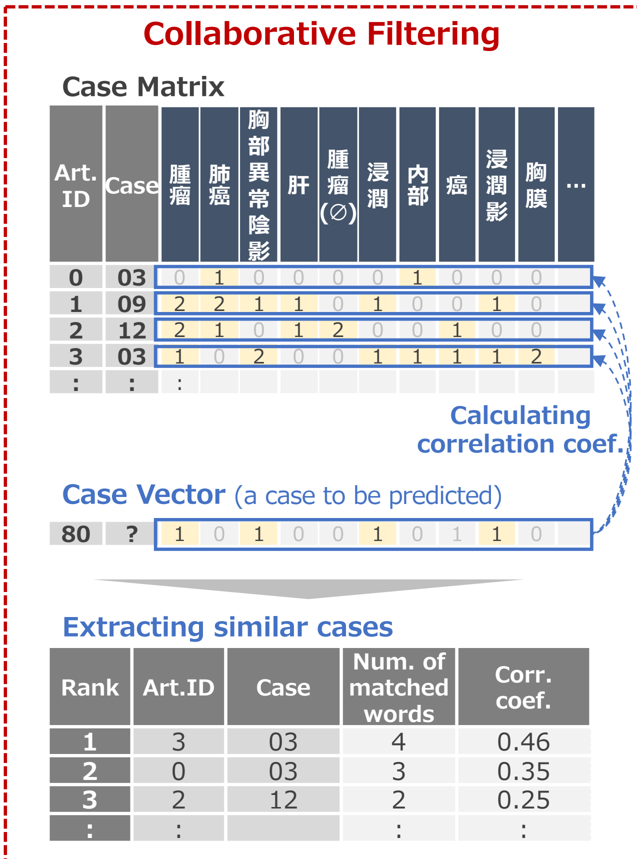

Figure 4: Collaborative filtering-based case identification framework for "case matrix"

4 EXPERIMENTS

4.1 Initial model development and prediction

A case matrix was extracted from training data based on the steps described in Figures 1 and 2, which had 72 cases in rows and 782 words in columns. The columns The case number of each article was masked and predicted from the case matrix of the other articles using user-based collaborative filtering. For training data, a matching rate between the predicted case numbers and true ones was 60% (43 of 71). There was no significant difference between correlation coefficient and other similarity indexes such as cosine similarity and Hamming distance.

In contrast, for test data, the evaluation results notified from the organizer was 0.3569.

4.2 Model improvement by dimensionality reduction

The high dimensionality of the case matrix with 782 columns seemed to make similar case identification difficult and negatively impact the matching rate. By Lasso regression [8], a kind of sparse modeling, which is a method of dimensionality reduction, the 782 words were reduced to 55 words that seemed to capture the features of each case essentially. Using the dimensionally-reduced case matrix, the matching rate increased to 72% (51 of 71) for training data.

5 CONCLUSIONS

The case matrix and collaborative filtering-based similar case identification method also successfully worked for radiology reports. This approach could be applied to other medical texts, such as pathology reports and medical progress notes, as a versatile method. However, there is still room for improvement in the matching rate for training and test data.

In general, a collaborative filtering approach has a "cold start problem" [9], which makes similar case identification difficult for a few cases. Some solutions have been proposed, such as clustering or semi-supervised learning [10].

In addition, the high dimensionality and sparseness of the case matrix possibly made case search difficult and reduced the matching rate. In fact, dimensionality reduction by sparse modeling was effective in improving the matching rate. This problem could also be resolved by a model-based collaborative filtering method such as non-negative matrix factorization [11-12].

In the future, I will consider the combination with model-based methods to improve accuracy and enhance functions such as the prediction of symptoms and complications.

REFERENCES

[1] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, Eiji Aramaki, Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task, *In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*. 2022.

[2] Noguchi R, Torikai K, Saito Y. Development of Discriminative Model for Disease Diagnostic Support and Effect of Orthographic Normalization Using Structured EMRs "Case Matrix". *In Proceedings of the 24th Spring Meeting of JAMI.* 2020

[3] Noguchi R, Torikai K, Saito Y. Collaborative Filtering-based Similar Case Matching by "Case Matrix" Developed from Text Data in Electronic Medical Records. *Japan Journal of Medical Informatics*. 2021;41(Suppl.):928-931

[4] Bronselaer G, De Tre. Concept‐relational text clustering. *Journal of intelligent systems*. 2012;27:970-93.

[5] Ito K, Nagai H, Okahisa T, et al. J-Medic: A Japanese disease name dictionary based on real clinical usage. *In Proceedings of the 11th International Conference on Language Resources and Evaluation.* 2018;2365-69.

[6] Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web*. 2001;285–295.

[7] Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*. 2003;7(1):76-80.

[8] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267-288.
[9] Ricci F, Rokach K, Shapira B. Recommender Systems Handbook Second Edition. *Springer*. 2015.
[10] Zhang M, Tang J, Zhang X, Xue X. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. *In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR '14)*. 2014:73-82.
[11] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401:788–91.
[12] Xin L, Mengchu Z, Yunni X, Qingsheng Z. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. *IEEE Transactions on Industrial Informatics*. 2014;10(2):1273-84.