Team-Nikon at NTCIR-12 MedNLPDoc Task

Hiroko Kobayashi
Nikon Corporation
Hiroko.Kobayashi@nikon.com

Toyoharu Sasaki
Nikon Corporation
Toyoharu.Sasaki@nikon.com

Toru Fujii
Nikon Corporation
Toru.Fujii@nikon.com

ABSTRACT

The phenotyping task of the NTCIR-12 MedNLPDoc Task [1] is a multi-labeling task retrieved from Japanese medical records. The team-Nikon participated in this task and proposed a new method that assigns the ICD codes by using Information Retrieval (IR) and reduces the magnitude of mistaken coding by using machine learning. When evaluated on development set, our system achieved F-scores of 29.2% and showed an effect of less mistaken coding compared to the IR method. On the other hand, in the test set, the effect of IR method is higher than the combined method of IR and machine learning.

Team Name

Team-Nikon

Subtask

Task-1: Phenotyping task

Keywords

Medical records, named entity recognition (NER), information retrieval, machine learning, logistic regression, conditional random field (CRF)

1. INTRODUCTION

There are various natural language processing (NLP) 'shared tasks (contests, competitions, challenge evaluations, critical assessments)' to encourage research on medical information retrieval. For example, the Informatics for Integrating Biology and Bedside (i2b2) [2] by the National Institutes of Health (NIH) is the best-known medical-related shared task and the ShARe/CLEF eHealth Evaluation Lab [3] is an organized European medical shared task. NTCIR-10 MedNLP task [4] and NTCIR-11 MedNLP 2 task [5] are shared tasks, evaluating the technologies that retrieve necessary information from medical reports written in Japanese. These tasks include three sub tasks, namely, entity removal task (de-identification task), medical term extraction task (complaint and diagnosis), and normalization task (ICD coding task). We participated in task-1: phenotyping task of MedNLPDoc in NTCIR12.The goal of this task is to assign International Codes for Diseases (ICD) from the text in medical records. ICD is the standard diagnostic coding system, which is maintained by the World Health Organization (WHO). The latest version of ICD is ICD-10,which consists of codes with first alphabet and several numbers. The developed system in this task can directly support an actual application for daily clinical services and can be used in many other areas of clinical studies.

In this paper, we have described the related works in Section 2, and have proposed our method in Section 3. We have presented our results in Section 4 and the conclusion of the study in Section 5.

2. RELATED WORKS

This task is considered as a multi-labeling task because of the its goal is to assign ICD-codes against per medical record.

Against multi-label problem, Berger [6] proposed a method that uses the convolutional neural network (CNN) and recurrent network with a gated recurrent unit (GRU) for higher performance than other methods. Recently, CNN and GRU have gained attention in machine learning field, which uses two million documents and 1000 potential labels of BioASQ Challenge [7] as training data.

On the other hand, the training corpus provided by the NTCIR-12 MedNLPDoc task consists of 200 individual medical records and 552 code types, and the average number of codes per record is 3.86. Thus, the difference in the training data size between the method proposed by Berger [6] and the NTCIR-12 MedNLPDoc task is in four digits. The data size of both NTCIR-12 MedNLPDoc task and NTCIR-11 MedNLP2 [5] task-2 are similar. The task of adding ICD-10 code to the medical term (training documents: 102, ICD-code mentions (<c>tag):3394 [8]) is similar to the task in which we participated. In the study by Fujino et al. [9] , the proposed method was used for increasing the training data, which included not only the annotated medical document sets provided by the organizers but also the dictionary pairs of medical terms and ICD codes. Moreover, the one-vs-rest approach with logistic regression was employed, and many other IR methods were proposed. The IR methods included the search of the nouns as queries by using the dictionary of medical terms and ICD-codes. For example, the method proposed by Kikui et al. [10].The case of medical records includes the cases of extract match, assigned ICD-codes, and the cases of a partial match. The filter is applied by the using the features of medical terms(prefix/suffix and construction). Fujino et al. [9] and Kikui et al. [10], focused on assigning the ICD-codes and not on reducing the misjudged assignments.

NTCIR-11 MedNLP2 task-2's data have already been labeled as complaints and diagnosis region in the medical records, whereas the task's goal is to assign the ICD-10 codes against the region. In contrast, the goal of this NTCIR-12 MedNLPDoc task is to assign the ICD-codes against the medical records. Therefore, the misjudgment in MedNLPDoc task risk is higher than that in NTCIR-11 task-2.

3. PROPOSED METHOD

In this task, a training data set of medical records is taken from "ICD Coding Training, Second Edition", written in Japanese for training Health Information Managers (HIMs) [11], and the ICD-codes are assigned based on the coding policy against the medical records.

3.1 Coding Example

There are few cases that are not coded under ICD-codes, despite the medical records including the medical terms (Figure 1, 2). For this study, there are two supposed cases: (Case-1) the medical record includes the negative expression about medical terms, and (Case-2) not code based on coding policy.

(a)Input

```
<data id="26" sex="m" age="38">
<text>
３カ月前から強い全身倦怠感とともに眼瞼部を中心に紅斑、顔の皮疹が出現する。発熱は認めなかったが症状の消長を繰り返し、次第に脱力感も加わり軽快しないため精査目的で入院となる。顔面の皮疹、脱力感、筋力低下、ゴットロン徴候などの症状から皮膚筋炎を疑い皮膚生検、筋生検を行った結果、筋線維の変性壊死および筋線維大小不同がみられ、皮膚筋炎の確定診断を得た。
・・・
</text>
```

(b)Output

```
<icd code="M331">皮膚筋炎性間質性肺炎</icd>
<icd code="R749">血清酵素値異常</icd>
```

Figure 1: Medical record sample(Case 1)

(a)Input

```
<data id="66" sex="m" age="45">
<text type="既往歴">なし
２００５年１月　１月初旬から咳が続き売薬購入するも改善なし.
２月３日　他院受診.
喀痰からＧ６号検出.
２５日　肺結核の診断にて当院紹介入院.
・・・
</text>
```

(b)Output

```
<icd code="A150">肺結核</icd>
```

Figure 2: Medical record sample(Case 2)

Figure 1 denotes a sample of the medical record (Case-1). This record does not code the ICD-code about the medical term ("発熱"), despite it being mentioned in the record. It is because the details include negative expression about the medical term ("発熱は認められなかったが"). Figure 2 denotes another sample of the medical record (Case-2) in which the medical record is written as "喀痰", but there is nothing coded under ICD-codes about "喀痰".As per the coding policy written by Aramaki et al. [1] , we should only code diseases or treatments, which are observed in a medical facility or where the coder belongs. "喀痰" is found in the other hospital ("他院").

3.2 System Overview

An overview diagram of our system is shown in Figure 3.Firstly, the system retrieve medical terms in the medical record as queries by using ICD dictionary, which consists of pairs of medical terms and ICD codes (assignment). However, the system excludes more likely errors ICD-codes by using machine learning (i.e., filtering).

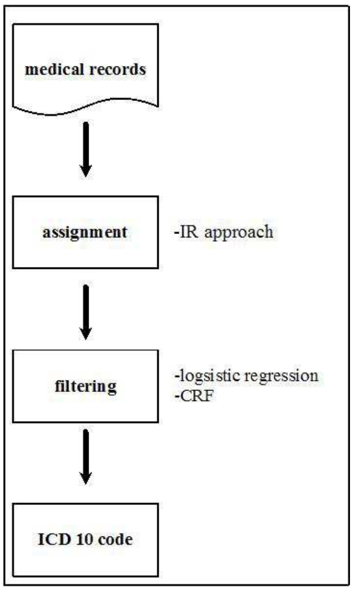

Figure 3: Method design

3.3 IR approach

To assign ICD-codes, we first excluded the details of the family histories from the medical records, followed by replacing the abbreviation of medical records with the replaced medical terms by using the dictionary that consists of pairs of abbreviations and medical terms in Wikipedia[1]. The system then retrieves the medical terms in the medical record and assigns ICD codes by using the dictionary consisting of medical terms and ICD codes (MEDIS Standard Medicine Masters[2] and MEDIS of synonyms and provided by the annotated data of NTCIR11 MedNLP task [5]) . The prefix string or suffix string of medical terms were deleted to prevent the retrieval omission and execute the partial matching. Sometimes, terms retrieved from the medical records include common strings, for example, "糖尿病" and "糖尿".In this case, we were assigned to extract a longer word, preferentially similar to the term extracted in the study by Nomura et al. [12].

[1] https://ja.wikipedia.org/wiki/

[2] https://www2.medis.or.jp/stdcd/byomei/index.html

3.4 Filtering

In this task, we have applied two kinds of filtering to work on the sample medical records described in section 3.1 (Case-1 and Case-2). They are (a) focus on modality in the sentence, and (b) focus on the misjudgment in coding in the training data.

3.4.1 Filtering-1 (modality detection)

In this section, we detected some modality expression for filtering as sequential labeling problem in the medical records. Conditional random field (CRF) is a type of statistical modeling method, which is used in sequential labeling problem and various other NLP tasks [13] [14]. In this study, we used CRF++[3] distribution as a tool of CRF, NTCIR 11 [5] and GSK Dummy Electronic Health Record Text Data (GSK2012-D)[4] as corpuses. In these corpuses, the symptom and diagnosis related expressions were marked as <c><\c>(c-tag), and the words and phrases were suggested as the modalities of the symptoms. We used two kinds of modality, i.e., symptoms that are not recognized (<c modality="Negation">) and disease of patients' family members (<c modality="Family">). To utilize the CRF, we converted the texts into word in IOB2 representation format and used Mecab[5], one of the major Japanese morphological analyzer, to obtain morphological features. The morphological features included parts of speech, inflected forms of word, and script types (Hiragana, Katakana, Chinese characters, or symbols). If the medical term recognized the modalities in medical records by using morphological classifications, then we did not assign ICD-codes.

3.4.2 Filtering-2 (coding possibility)

To extract the cause the discriminating error in the ICD-code, we first executed the IR approach of section 3.3 against the training data ('*MedNLPDoc_TRAIN_v5.xml*') that was provided by the organizers. The case assigned ICD-code is not the annotated one from the training data, hence, we calculated the errors in the ICD-code (c_1,c_2,…c_l). These codes defined that the error appeared over five times in the training data set. Moreover, we applied logistic regression model to classify these codes. The logistic regression model is a type of method that predicts one of dependent variable (0 or 1) from one or more independent variables [15]. We used open source software LIBLINEAR[6] and medical records assigned by IR approach in training data. In the case of correctly assigned record annotation is 1 about c_i, otherwise the annotation is 0. Moreover, to extract the features from medical records, the first noun words were analyzed by using Mecab[7]. After excluding the stop words from the records, the important words were extracted from the words by calculating TF-IDF (Term Frequency-Document Frequency) weights.

4. RESULTS

We evaluated the proposed our method between the development set and the test set provided by the organizers (Table 1, 2, 3, 4). The development set from the training data set is divided into data set for training and evaluating. Nearly 20% of the all the data set were used as evaluating data(Table 1). The test data consisted of 78 clinical texts, and 3 professional human coders added codes. The case in which all the coders added the codes is considered as 'SURE' code; when two or three coders added codes is considered as 'MAJOR' code; and when at least one coder added codes is the 'POSSIBLE' code (Tables 2, 3, 4).

We tried three method, Method-1:IR approach,Method-2: IR approach+filtering-2, and Method-3:IR+filtering-1 + filtering-2.The performance of this phenotyping task was assessed using the F-score (β =1), precision, and recall **[16]**. Precision is the percentage of correct codes found in the participant's system ; recall is the percentage of codes present in the corpus that were found by the system; and F-score is the harmonic mean of precision and recall.

development set	**precision**	**recall**	**F-score**
IR	23.7%	29.5%	26.3%
IR +filteing-2	30.7%	27.7%	29.1%
IR + filtering-1+ filtering-2	30.9%	27.7%	29.2%

Table 1: Result (Development set)

test data (SURE)	**precision**	**recall**	**F- score**
IR	22.3%	47.0%	30.3%
IR +filteing-2	26.7%	25.6%	26.1%
IR + filtering-1+ filtering-2	26.5%	25.3%	25.9%

Table 2: Result (Test Data: SURE)

test data (MAJOR)	**precision**	**recall**	**F- score**
IR	40.2%	48.7%	44.0%
IR +filteing-2	39.3%	23.2%	29.1%
IR+ filtering-1+ filtering-2	39.1%	22.9%	28.9%

Table 3: Result (Test Data: MAJOR)

test data (POSSIBLE)	**precision**	**recall**	**F- score**
IR	48.0%	31.8%	38.2%
IR +filteing-2	48.4%	15.0%	22.9%
IR+filtering-1+ filtering-2	48.3%	14.9%	22.8%

Table 4: Result (Test Data: POSSIBLE)

The effect of filtering of the development data set in Method-2 and Method-3 was found to be higher than that in Method-1. On the other hand, the effect of using the test set was not shown. The recall of test data set in Method-1 is the highest; therefore, correct assignment data was excluded from the data set provided for filtering.

5. CONCLUSION

In this paper, we outlined the methods used for obtaining our experimental results for the team-Nikon and discussed the derived results. In this development set, we showed the effect of less misjudged coding. Contrastingly, we did not show the effect in the test set. The performance in this provided method may be improved by analyzing the incorrect code in the test data.

[3] https://taku910.github.io/crfpp/

[4] http://www.gsk.or.jp/catalog/gsk2012-d/

[5] http://taku910.github.io/mecab/

[6] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

[7] http://taku910.github.io/mecab/

6. REFERENCES

[1] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. 2016. Overview of the NTCIR-12 MedNLP Doc Task, In *Proceedings of NTCIR-12*.

[2] Ozlem, U. 2008. Second i2b2 workshop on natural language processing challenges for clinical records, in *AMIA Annual Symposium proceedings*. 1252-1253.

[3] ShARe/CLEF eHealth Evaluation Lab. 2013 [cited 2014/06/04; Available from: https://sites.google.com/site/shareclefehealth/.

[4] Morita, M., Kano, Y., Ohkuma, T., Miyabe M., and Aramaki, E. 2013. Overview of the NTCIR-10 MedNLP task, *In Proceedings of NTCIR-10*.

[5] Aramaki, E., Morita, M., Kano, Y.,and Ohkuma, T., 2014. Overview of the NTCIR-11 MedNLP-2 Task, *In Proceedings of NTCIR-11*.

[6] Berger, M. J., 2015. Large Scale Multi-label Text Classification with Semantic Word Vectors.

[7] Tsatsarnois, G., et al. 2015.An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition.*BMC bioinformatics.*

[8] Chen, S., Lai, P., Tsai,Y., Chung,J., Hsiao, S,. and Tsai,R., 2014,NCU IISR System for NTCIR-11 MedNLP-2 Task , *In Proceedings of NTCIR-11*.

[9] Fujino, A., Suzuki, J., Hirao,T., Kurasawa, H., and Hayashi, K., 2014,SCT-D3 at the NTCIR-11 MedNLP-2 Task, *In Proceedings of NTCIR-11*.

[10] Kikui, G,.and Tajima,Y., 2014,The OKPU System in NTCIR11 MedNLP2:An IR Approarch to ICD-10 Code Identification, *In Proceedings of NTCIR-11*.

[11] 鳥羽克子,ICD コーディングトレーニング(編集),診療情報管理東京ネットワーク(編集),医学書院

[12] Nomura, Y., Suenaga, T., Satoh, D., Ohki, M., and Takai,T., 2013,Medical Information Extracting System by Bootstrapping of NTTDATA at NTCIR-10 MedNLP Task, *In Proceedings of NTCIR-10*.

[13] Nakamura, T., Kudo, K., Shikata, S., Miyabe, M,.and Aramaki, E.,2014, kyoto:Kyoto University Baseline at the NTCIR-11 MedNLP-2 Task, *In Proceedings of NTCIR-11*.

[14] Tawara, Y., Omura, M.,and, Miura, M., 2014,Incorporating Unsupervised Features into CRF based Named Entity Recognition, *In Proceedings of NTCIR-11*.

[15] 春野瑞季,村上仁一,徳久雅人,ロジスティック回帰分析を用いたパターンに基づく統計翻訳,2015,言語処理学会

[16] van Rijsbergen, C.J. 1975. *Information Retrieval.* Buttersworth,London.