MedNLPDoc: Japanese Shared Task for Clinical NLP

Eiji Aramaki
NAIST
aramaki@is.naist.jp

Yoshinobu Kano
Shizuoka University
kano@inf.shizuoka.ac.jp

Tomoko Ohkuma
NAIST
tomokoohkuma@gmail.com

Mizuki Morita
Okayama University
mizuki@okayama-u.ac.jp

Abstract

Due to the recent replacements of physical documents with electronic medical records (EMR), the importance of information processing in medical fields has been increased. We have been organizing the MedNLP task series in NTCIR-10 and 11. These workshops were the first shared tasks which attempt to evaluate technologies that retrieve important information from medical reports written in Japanese. In this report, we describe the NTCIR-12 MedNLPDoc task which is designed for more advanced and practical use for the medical fields. This task is considered as a multi-labeling task to a patient record. This report presents results of the shared task, discusses and illustrates remained issues in the medical natural language processing field.

1 Introduction

Medical reports using electronic media are now replacing those of paper media. Correspondingly, the information processing techniques in medical fields have radically increased their importance. Nevertheless, the information and communication technologies (ICT) in medical fields tend to be underdeveloped compared to the other fields [1].

Processing large amounts of medical reports and obtaining knowledge from them may assist precise and timely treatments. Our goal is to promote developing practical tools that support medical decisions. In order to achieve this goal, we have been organizing 'shared tasks (contests, competitions, challenge evaluations, critical assessments)' to encourage research in medical information retrieval. Among the various shared tasks, one of the best-known medical-related shared tasks is the Informatics for Integrating Biology and the Bedside (i2b2) by the National Institutes of Health (NIH), which started in 2006 [2]. The Text Retrieval Conference (TREC), which addresses more diverse issues, also launched the Medical Reports Track [3]. Shortly after the NTCIR-10 MedNLP task, the first European medical shared task, the ShARe/CLEF eHealth Evaluation Lab [4], was organized. This shared task focuses on natural language processing (NLP) and information retrieval (IR) for clinical care. While they are targeted only at English texts, medical reports are written in native languages in most countries. Therefore, information retrieval techniques in individual language are required to be developed.

We organized the NTCIR-10 and NTCIR-11 MedNLP tasks (shortly MedNLP) [5] which were the first and second shared tasks, evaluating technologies that retrieve important information from medical reports written in Japanese. These previous tasks include three sub tasks: named entity removal task (de-identification task), disease name extraction task (complaint and diagnosis), and normalization task (ICD coding task). These tasks correspond to elemental technologies for computational systems which support diverse medical services.

Following the success of these MedNLP tasks, we designed the NTCIR-12 MedNLPDoc task to be more advanced and practical. In this MedNLPDoc task, we provided a new challenging task where participants' systems infer disease names in ICD (International Codes for Diseases) from textual medical

Proceedings of the Clinical Natural Language Processing Workshop,
pages 13–16, Osaka, Japan, December 11-17 2016.

records. Due to this practical setting, task participants' systems could directly support an actual daily clinical services and clinical studies in various areas.

2 Task & Corpus

2.1 ICD Code

The International Classification of Diseases (ICD) is the standard diagnostic coding system used in many countries for epidemiology, health management and clinical purposes. It is used to monitor the incidence and prevalence of diseases and other health problems, proving a picture of the general health situation of countries and populations. In the latest version of the ICD coding system, ICD-10, each ICD code consists of a single alphabet prefix and two digits of numbers. In addition to these three characters that represents a major classification, more detailed classification can be represented by several digits of additional numbers as suffix, up to six characters in total. Because the major categories are limited to 21 sections, the major categories include a set of similar diseases.

2.2 ICD Coding Task

We provided a training data set of medical records that is taken from "ICD Coding Training, Second Edition", written in Japanese for training Health Information Managers (HIMs). We organized the phenotyping task, in which the participants are required to assign ICD-10 code(s) to a given medical record.

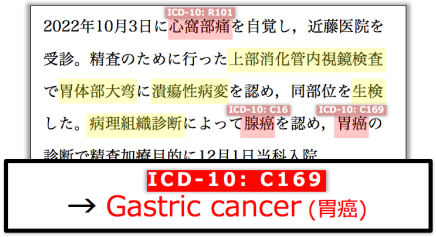

Figure 1: coding task: The participants will assign ICD-10 codes from texts. This example should be assigned as C169, which means Gastric cancer. In this case, only one ICD-10 code was assigned, but in general, one or more codes should be assigned.

Table 1: Number of Code Assigned.

	# (valiance)	Min.	Max.
SURE	2.8 (±1.6)	0	8
MAJOR	5.2 (±2.0)	1	11
POSSIBLE	9.4 (±3.8)	2	19

2.3 Corpus

We created a medical record corpus for this task which includes 278 individual medical records (200 were used for training, and the other 78 used for test). The average number of sentences per record is 7.82. The average number of codes per record is 3.86 (total 1073 codes and 552 variants). In test set, three professional human coders (more than one-year experience) individually added ICD-10 codes. We defined three different code sets as follows.

- **SURE (S)**: sure code set consists of codes that all coders (three persons) utilized.
- **MAJOR (M)**: major code set consists of codes that two or three coders utilized.
- **POSSIBLE (P)**: possible code set consists of codes that at least one coder utilized.

We derived three types of gold standard data for each code set above. Note that there is a relationship of $S \in M \in P$ (SURE is a subset of MAJOR, MAJOR is a subset of POSSIBLE). The inter annotation ratio (**IAA**) between three humans (Human-A, Human-B, and Human-C) is defined by the following formula:

$$\mathbf{IAA} = \Sigma_{d \in D} |A_d \cap B_d \cap C_d| \setminus |A_d \cup B_d \cup C_d| / |D|.$$

where D is the set of all records. $|D|$ is the number of records. $|A_d \cap B_d \cap C_d|$ is the number of SURE codes in the record *d*. $|A_d \cup B_d \cup C_d|$ is the number of POSSIBLE codes in the record *d*.

2.4 Evaluation

Performance of the coding task was assessed using the F-score (β=1), precision, and recall. Precision is the percentage of correct codes found by a participant's system. Recall is the percentage of codes presented in the corpus that were found by the system. F-score is the harmonic mean of precision and recall.

The three human coders were also evaluated by this measure. The average results are as follows: Av. Sure Precision=0.168, Av. Sure Recall=0.388, and Av. Sure F-measure=0.235.

3 Result

The participating systems are shown in Table 2. Roughly, the systems are classified into three types: (1) machine learning approach (team A, B, E, and G), (2) rule based approach (team C. D and H), and (3) their combination (team C).

3.1 Machine Learning V.S. Rule-based

The performance is shown in Figure 2. Among all systems, the highest performance system is provided by the SYSTEM-C in the SURE metrics. The system is based on heuristic rules, indicating that rule-based approaches still have its advantage. Considering machine learning approaches have been outperforming rule based approaches in most of the other NLP fields, this result is remarkable for future system designing in the medical domain.

In the other metrics (MAJOR and POSSIBLE), the system-G3 and the system E achieved better performance than the SYSTEM-C. Not like the SYSTEM-C, the SYSTEM-G3 fully implemented by the multiple machine learning methods. Also, the SYSTEM E system partly utilized machine learning, but it also employs rule-based features that represent coding heuristics.

In summary, the overall result indicates the advantages of traditional rule based approach. These results were caused by two reasons: (1) the corpus size of this task is relatively small than the other tasks, and (2) the classification space (the number of code) is huge. This result revealed that current machine learning techniques still suffer from such conditions.

3.2 Contribution of Extra Resources

Another viewpoint of this task is the contribution of extra resources. Almost all participants used the MEDIS Standard Masters (MDS) and some used other language resources. While this implies that a medical dictionary is the most useful tool to this task. The SYSTEM-D calculated similarity scores between medical vocabulary n-grams and word n-grams in EMR. The SYSTEM-H calculated edit-distances and used their scores as features of CRF. The SYSTEM-A used three dictionaries in addition to MDS. They used *Kuromoji* morphological analyzer with their customized dictionary. In summary, most of the teams have relied on the existing language resources, and its quality and quantity varies the team performance.

3.3 Strategy

The strategies of the systems are characterized by two parameters; (1) the average number of codes and (2) the variance of codes. Table 3 presents the average number of codes assigned by the high performance three systems (SYSTEM C, E, and G3). The SYSTEM-G3 assigns more codes rather than the others (high recall-oriented). In contrast, the SYSTEM-C ascend only 2.0 codes in average (high precision-oriented).

Another parameter is the distribution of codes. Figure 3 shows the distribution of codes of these systems. The SYSTEM-C handles a narrow coding spaces, in which the most of codes are assigned in Z**, R** or C**. This also indicates that the SYSTEM-C aims to obtain the high precision.

Table 2: participant system.

Team	Sources	Methods
A	ICD-10(en), Wikipedia, Google/Yandex MT, HUG(fr)	rule base
B	MDS, ICD-10	machine learning (CRF)/ Edit distance (as features)
C	MDS, Wikipedia	Rule based
D	MDS, ICD training book	string similarity measure
E	MDS	Rule based (as features), machine learning (CRF)
F	MDS, training data	search engine (using named entity based keywords?)
G	MDS	machine learning (CRF,LIBLINER (SVM))
H	MDS	NA (Exact Match)

* MDS indicates the ICD Dictionary, MEDIS Standard Masters.
* CRF indicates the conditional random fields.

Table 3: Number of Code Assigned.

SYSTEM	# of codes	Min.	Max.
C	2.0	0	7
E	3.4	1	8
G3	6.6	8	14

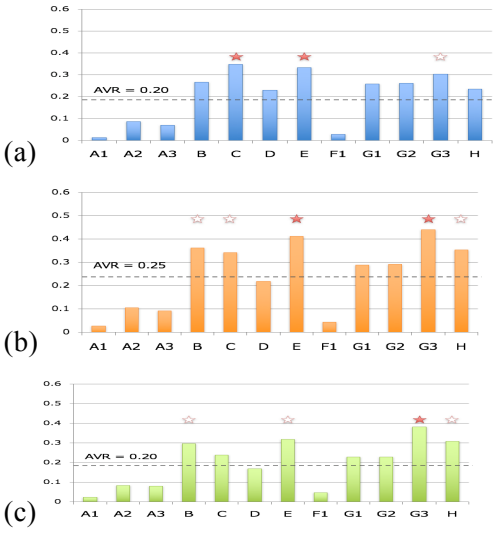

Figure 2: F-measure in SURE (a), MAJOR (b), and POSSIBLE (c).

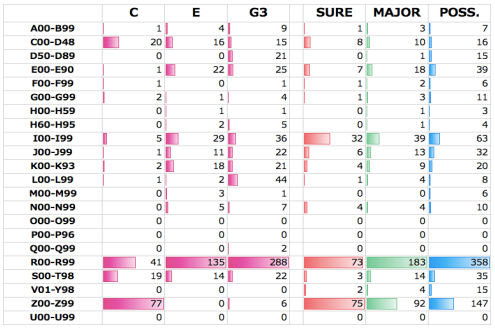

Figure 3: Code Distribution of the best three systems.

4 Conclusion

This paper describes the NTCIR-12 MedNLPDoc task which is a multi-labeling task, ICD-10 coding, to a patient record. This report presents results of the shared task, discusses and illustrates remained issues in the medical natural language processing field. Still, rule-based approaches have demonstrated the advantage in this task, requiring the future development of machine learning approaches that deal with small data.

Reference

[1] [r Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'Avolio, L.W., Savova, G.K., and Uzuner, O. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, 18, 540–543.

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

[3] Voorhees, E.M. and Hersh, W. 2012. Overview of the TREC 2012 Medical Records Track. in The Twentieth Text REtrieval Conference.

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

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