Overview of the NTCIR-12 MedNLPDoc Task

Eiji Aramaki
Kyoto University
eiji.aramaki@gmail.com

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

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

Tomoko Ohkuma
Fuji Xerox Co., Ltd.
ohkuma.tomoko@fujixerox.co.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.

Keywords

Medical records, electronic medical records (EMR), named entity recognition (NER), shared task and evaluation

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 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 & MATERIALS

2.1 What is 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. ICD is maintained by the World Health Organization (WHO) within the United Nations System.

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 MedNLP 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 following two subtasks:

- Phenotyping task: the participants are required to assign ICD-10 code(s) to a given medical record. This task corresponds to the so-called phenotyping task in the medical research field.
- Creative task: in this subtask, we welcomed participants' creative ideas that help us utilizing real world products. Especially, we expected a new task plan or annotation scheme for next MedNLPDoc-2.

2.3 Coding Policy

The followings are the major policies in our coding:

1) We only assign codes to diseases or treatments that are conducted in a medical facility where the coder belongs.

2) We assign codes to both medical histories related to the existing diseases and previous medical histories mentioned in the medical records.

3) We assign codes to not only diseases related to the primary disease but also diseases that are/were needed to be treated, even if they have no relation with the primary disease.

4) For diseases that include coexistence symptoms and complications, it needs to pay attention to the relation between the primary disease and additional diseases. For example, in the case of Pneumonia “anuresis“, a corresponding ICD-10 code should be assigned. However, in the case of Prostate, no code should be assigned because this is just a reference comment in the medical record.

2.4 Coding Example

Figure 1 presents several coding examples.

In the example 1, the primary disease, “肺結核“, has an ICD code “A150”. The other disease terms are out of coding target, e.g., “喀痰“ is ambiguous, and the disease is found in the other hospital (”他院“).

In the example 2, we could assume that this patient had amentia from the text “吐血” in the medical record. The decision of the grade of amentia depends on the situation of bleeding. In this case, the diagnostic “急性出血後貧血” and the corresponding ICD-10 code “D62” were added, and “K270” could be additionally coded. Like the example 2, some of medical records do not contain explicit names of diagnostics, but coders need to determine diseases and codes from medical histories and situations.

More information about the coding manner is available in the commentary parts of the ICD training book [6]. Detailed data format is shown in Figure 2.

(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.
ＩＮＨ　０．４，　ＲＦＰ　０．４，　ＳＭ　１ｇ／日，ＰＺＡ　１．５ｇ／日で化学療法スタート.
</text>
```

(b) Output

```
<icd code="A150">肺結核</icd>
```

(c) Input

```
<data id="68" sex="m" age="49">
<text>
２００４年１２月２～１６日，前回入院.
今回２回目の入院.
前回他院にてアメーバ肝膿瘍の手術予定であったが，術前の検査でＨＩＶ陽性であったため当院入院.
ＳＴＳ陽性.
今回上記の疾患について外来フォロー中であったが吐血で入院.
食道潰瘍が判明.
</text>
```

(d) Output

```
<icd code="K221">食道潰瘍</icd>
<icd code="D62">急性出血後貧血</icd>
<icd code="R75">ＨＩＶ陽性</icd>
<icd code="A530">ＳＴＳ陽性</icd>
<icd code="Z861">アメーバ肝膿瘍</icd>
<icd code="K270">吐血</icd>
```

Figure 1: Coding Example.

```
<?xml version="1.0" encoding="utf-8" ?>

<!DOCTYPE root[
  <!ELEMENT root (data+)>
  <!ELEMENT data (text+, icd+)>
  <!ELEMENT text (#PCDATA)>
  <!ELEMENT icd (#PCDATA)>
  <!ATTLIST data id CDATA #REQUIRED sex CDATA
#REQUIRED age CDATA #REQUIRED>
  <!ATTLIST text type cdrom (yes|no) "no">
  <!ATTLIST icd code CDATA #REQUIRED>
]>
```

Figure 2: Data format.

2.5 Corpus Statistics

We created a medical record corpus for this task which includes 200 individual medical records. The average number of sentences per record is 7.82. The average number of codes per record is 3.86. 552 code types appeared in the corpus. The inter-agreement ratio between annotators are presented in Section 3.

3. METHODS

3.1 Test Set data

Test data set consists of 78 clinical texts, which were randomly selected from the past State Examinations. Question sentences and graphics were eliminated from the original documents. Then, 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∈M∈P (SURE is a subset of MAJOR, MAJOR is a subset of POSSIBLE).

3.2 Evaluation method

Performance of the coding task was assessed using the F-score (β=1), precision, and recall [8]. 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.

We employed three matching levels as follows:

- LEVEL 4: Exact match.
- LEVEL 3: Partial match in the first three letters in the code.
- LEVEL 0: Category match, classified following blocks of codes, that define similar diseases and related health problems and consist of chapters of ICD-10 [9].

For example, codes "A169" and "A160" are considered as LEVEL 3 match. Evaluating in accordance with categories for similar diseases and related health problems, "C00" and "D48" are part of the chapter II of ICD-10 and are considered as LEVEL 0 match.

In total, we have three gold standard data sets, three matching methods and three matching levels. Therefore, three sets of precision, recall and F-measure and in total 27 scores are calculated. For example, SURE and LEVEL 3 match results consist of:

- Precision $^{LV3}_{sure}$ = |S∩R| / |R|
- Recall $^{LV3}_{sure}$ = |S∩R| / |S|
- F-measure = 2 ・ Precision $^{LV3}_{sure}$ ・ Recall $^{LV3}_{sure}$ / (Precision $^{LV3}_{sure}$ + Recall $^{LV3}_{sure}$)

3.3 Inter-agreement Ratio between Annotators

The inter-agreement ratio between annotates are shown as follows. Considering the annotators' skill, these value are small, indicating the difficulty of the ICD10 coding.

Sure Precision^{LV4}	0.168
Sure Recall ^{LV4}	0.388
Sure F-measure ^{LV4}e	0.235

4. RESULTS

4.1 Participating systems

The participating systems are shown in Table 1. 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). For the detailed of the system, see each system paper.

Table 1: 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.

4.2 Performances

The performance is shown in Table 2, consisting of (a) exact match, (b) rough match, and (c) category match.

Among all systems, the highest performance system is provided by the team "C", which shows the best performance in the half of all metrics (13/27 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.

Not like the top systems, the second rank system, "G3", fully implemented by the multiple machine learning methods. The system shows the best performance in the 12 metrics of all (12/27 metrics). The system utilized both of CRF based diseases name extraction, and the SVM based modality classification.

The third rank system, provided by the team "E", shows the best performance in one metrics. The team "E" system basically 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.

4.2 A Usage of medical dictionaries

The results of this task show that annotating ICD-code to EMR is promising. It is relatively easy to start NLP in medical domain rather than others because huge medical lexicons are already available, such as MEDIS Standard Masters (MDS).

Almost all participants used the MDS and some used other language resources. While this implies that a medical dictionary is the most useful tool to this task, usage of the language resources varies with team.

Baseline system employed exact match in the simplest way. [NIKON], [UE] and [NIL] also used exact match.

[Matsu] calculated similarity scores between medical vocabulary n-grams and word n-grams in EMR. [HCU] calculated edit-distances and used their scores as features of CRF.

[KIS] 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.

5. 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.

6. REFERENCES

[1] 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*.

[6] 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

[7] Japanese Society of Internal Medicine. 2014. [cited 2014 2014/06/04]; Available from: http://www.naika.or.jp/.

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

[9] International Statistical Classification of Diseases and Related Health Problems 10th Revision 2010

[Available from: http://apps.who.int/classifications/icd10/brow

Table 2: Overall results of (a) exact match, (b) rough match, and (c) category match.

(a) Exact match

	SURE			MAJOR			POSSIBLE		
Team	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**
A1	0.02	0.010	0.013	0.033	0.021	0.026	0.045	0.016	0.023
A2	0.087	0.087	0.087	0.132	0.087	0.105	0.151	0.057	0.083
A3	0.058	0.087	0.07	0.092	0.093	0.092	0.11	0.063	0.080
B	0.209	0.364	0.266	0.361	0.363	0.362	0.42	0.23	0.297
C	0.423	0.295	0.348	0.597	0.239	0.341	0.681	0.145	0.239
D	0.237	0.223	0.23	0.313	0.168	0.219	0.374	0.109	0.169
E	0.316	0.353	0.334	0.524	0.338	0.411	0.6	0.217	0.319
F1	0.018	0.064	0.028	0.032	0.072	0.044	0.044	0.05	0.047
F2	0.065	0.042	0.051	0.096	0.044	0.06	0.166	0.038	0.062
F3	0.086	0.040	0.054	0.12	0.039	0.058	0.199	0.032	0.055
G1	0.265	0.253	0.259	0.391	0.229	0.289	0.483	0.149	0.228
G2	0.267	0.256	0.261	0.393	0.232	0.291	0.484	0.15	0.229
G3	0.223	0.470	0.303	0.402	0.487	0.44	0.48	0.318	0.382
H	0.173	0.388	0.235	0.314	0.408	0.354	0.37	0.265	0.309

(b) Rough match

	SURE			MAJOR			POSSIBLE		
Team	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**
A1	0.066	0.044	0.053	0.132	0.07	0.091	0.159	0.057	0.083
A2	0.169	0.139	0.153	0.269	0.156	0.198	0.324	0.123	0.178
A3	0.11	0.146	0.126	0.205	0.176	0.189	0.242	0.137	0.175
B	0.221	0.383	0.28	0.39	0.399	0.394	0.473	0.286	0.356
C	0.47	0.317	0.379	0.646	0.261	0.371	0.729	0.167	0.271
D	0.251	0.226	0.238	0.333	0.176	0.23	0.41	0.13	0.197
E	0.336	0.377	0.355	0.553	0.36	0.436	0.676	0.266	0.382
F1	0.048	0.165	0.074	0.093	0.157	0.117	0.122	0.12	0.121
F2	0.125	0.075	0.094	0.212	0.089	0.125	0.289	0.064	0.104
F3	0.145	0.072	0.096	0.229	0.08	0.119	0.303	0.053	0.091
G1	0.268	0.264	0.266	0.415	0.253	0.315	0.515	0.173	0.259
G2	0.27	0.267	0.268	0.416	0.256	0.317	0.516	0.174	0.26
G3	0.232	0.497	0.316	0.423	0.517	0.465	0.533	0.382	0.445
H	0.184	0.414	0.251	0.338	0.438	0.382	0.414	0.323	0.363

(c) Category match

	SURE			MAJOR			POSSIBLE		
Team	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**	**Precision**	**Recall**	**F**
A1	0.165	0.179	0.172	0.266	0.207	0.233	0.356	0.356	0.356
A2	0.311	0.241	0.272	0.474	0.292	0.361	0.612	0.612	0.612
A3	0.231	0.305	0.263	0.359	0.367	0.363	0.46	0.46	0.46
B	0.407	0.557	0.47	0.577	0.578	0.578	0.673	0.673	0.673
C	0.662	0.504	0.572	0.809	0.441	0.571	0.854	0.854	0.854
D	0.464	0.414	0.437	0.597	0.407	0.484	0.668	0.668	0.668
E	0.542	0.533	0.537	0.741	0.548	0.63	0.848	0.848	0.848
F1	0.21	0.525	0.3	0.305	0.545	0.391	0.41	0.41	0.41
F2	0.34	0.218	0.266	0.53	0.251	0.341	0.652	0.652	0.652
F3	0.362	0.215	0.27	0.558	0.245	0.34	0.66	0.66	0.66
G1	0.467	0.507	0.486	0.624	0.513	0.563	0.722	0.722	0.722
G2	0.467	0.507	0.486	0.624	0.513	0.563	0.722	0.722	0.722
G3	0.42	0.676	0.518	0.601	0.693	0.644	0.712	0.712	0.712
H	0.398	0.605	0.48	0.575	0.64	0.606	0.677	0.677	0.677

- Team names are masked.
- Underlined value is the highest score in each metrics.