Overview of the NTCIR-11 MedNLP-2 Task

Eiji Aramaki
Kyoto University
eiji.aramaki@gmail.com

Mizuki Morita
The University of Tokyo
mizuki@sict.i.u-tokyo.ac.jp

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

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

ABSTRACT

Electronic medical records are now often replacing paper documents, and thus the importance of information processing in medical fields has increased. We have already organized the NTCIR-10 MedNLP pilot task. It has been the very first shared task attempt to evaluate technologies to retrieve important information from medical reports written in Japanese, whereas the

NTCIR-11 MedNLP-2 task has been designed for more advanced and practical use for the medical fields. This task was consisted of three sub tasks: (Task 1) the task to extract disease names and dates, (Task 2) the task to add ICD-10 code to disease names, (Task 3) free task. Ten groups (24 systems) participated in Task 1, 9 groups (19 systems) participated in Task 2, and 2 groups (2systems) participated in Task 3. This report is to present results of these groups with discussions that are to clarify the issues to be resolved in medical natural language processing fields.

Keywords

Medical records, electronic health records (EHR), 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 to support medical decisions. In order to achieve this goal, we have organized '*shared tasks (contests, competitions, challenge evaluations, critical assessments)*' to encourage research in 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), started in 2006 [2]. The Text Retrieval Conference (TREC), which addresses more diverse issue, also launched the Medical Reports Track [3]. Shortly after out the NTCIR-10 MedNLP task was organized, the first European task was also organized. It was the ShARe/CLEF eHealth Evaluation Lab [4], and this shared task focusing on natural language processing (NLP) and information retrieval (IR) for clinical care. However, they are targeted only at English texts. On the contrary,

Table 1. The Clinical field distribution of reports.

	Train		Dry		Test	
	D	Q	D	Q	D	Q
Disorder of the Alimentary Tract	4 (4)	8	1	0	3	3
Liver, Biliary Tract & Pancreas	2 (2)	7	0	0	0	3
Cardiovascular System	10 (12)	7	0	0	3	3
Endocrinology, Metabolism & Nutrition	7 (5)	6	0	0	1	3
Disorders of the Kidney & Urinary Tract	2 (4)	6	0	0	3	3
Immune System & Immune-Mediated Injury	5 (5)	4	0	0	6	2
Disorders of the Hematopoietic System	2 (1)	5	0	0	4	2
Infectious Disease	4 (6)	6	0	0	2	3
Disorders of the Respiratory System	11 (11)	8	1	1	3	2
	47 (50)	57	2	1	25	24

* The number in a bracket is the number of reports in MedNLP-1. D indicates D-rep. Q indicates Q-rep.

Medical reports are written in native languages in most countries. Therefore, information retrieval techniques in each language should be developed.

The NTCIR-10 MedNLP pilot Task (shortly MedNLP-1) [5] was the first shared task attempt to evaluate technologies to retrieve important information from medical reports written in Japanese. In this task, the test set was consisted of 50 medical records. Using this dataset, we designed three sub tasks: (Task 1) the task to remove the named entity (de-identification task), (Task 2) the task to extract disease names (complaint and diagnosis), and (Task 3) free task (participants design their original tasks). These tasks represent elemental technologies that are used to develop computational systems supporting widely diverse medical services. Development has yielded 22 systems for Task 1, 15 systems for Task 2, and 1 system for Task 3.

Following the success of MedNLP-1, the NTCIR-11 MedNLP-2 task is designed for more advanced and practical for the medical fields. In this task, the test set is consisted of 49 medical records. The task is consisted of three sub tasks: (Task 1) the task to extract disease names and time date (Task 2) the task to add ICD-10 code to disease names, (Task 3) free task. Note that Task 1 is

similar to Task 2 of MedNLP-1, but Task 2 of NTCIR-11 is a new task. This term normalization process is required to some medical applications such as information retrieval, data mining and so on. Task 2 in MedNLP-2 is a step forward to the next stage for practical usages. Ten groups (24 systems) participated in Task 1, 9 groups (19 systems) participated in Task 2, and 2 groups (2 systems) participated in Task 3 for MedNLP-2.

2. MATERIALS

2.1 Corpus

The material of MedNLP-2 contained two types of data: (1) the Dummy Patients' Medical Reports (shortly, **D-Rep**), and (2) the Questions from the past State Examinations extracted from the actual past state examinations (**Q-Rep**). The question sentences and graphics were eliminated. The clinical distribution of the reports is shown in Table 1; as shown, the material covers most clinical fields.

D-rep was constructed from *'dummy'* medical reports that doctors had written for their *'dummy'* patients. Since medical reports usually include extremely sensitive personal information about patients and others, such as patients' families, friends, and colleagues, it was difficult to perfectly remove such personal information. Therefore, the physicians had exclusively created medical reports of putative or imaginary patients in Japanese for this study. Each medical report typically contained the chief complaint, patient's disease history, diagnosis, treatments, clinical course, and the outcome.

In order to investigate the consistency of the dummy records, we had examined a blind check using the mixture of the dummy reports and the real reports. First, we randomly picked 10 records from the dummy reports. We also picked 10 records gathered by the Japanese Society of Internal Medicine [6]. Five evaluators (two medical staffs, and three non-medical staffs including one of authors) had classified dummy records from the mix of two types of records. The result is shown in Table 2. Although the chance level is 50% (=10/20), the results were close to the chance level, supporting the validness of the dummy data. The differences between medical staffs and the other staffs were small; also implying that the dummy data was equivalent to the medical viewpoint.

2.2 Annotation

2.2.1 Date time

Date and time related expressions were marked with **<t></t>** (t-tag) as follows.

Compound noun: time related nouns and phrases were marked. If a compound noun contained terms that were not related to tense or any other time related information, such terms were excluded from the tag.

Non-numerical expression relating tense: numerical expressions that indicated date and/or time were marked. Non-numerical expressions were not marked, unless they indicated specific dates and/or time.

Relative time: both '*absolute*' and '*relative*' time/tense expressions were marked.

Duration: the whole clauses and phrases of time expression including punctuations and hyphenations (e.g. "—", "～") were marked together.

Table 2. Accuracy of the dummy classification.

Evaluator	Accuracy
Medical (physician) (n=2)	60.0%
Non medical (n=3)	56.3%

Table 3. Modality types.

Type	Description
Positive	Actually recognized symptoms. (Default)
Negation	Symptoms that are NOT recognized.
Suspicion	Suspected and unconfirmed diseases.
Family	Diseases of the patients' family members.

Non-clinical expression: expressions that were not directly related to medical information were not marked.

Case marker: Japanese case markers [Jo-shi], such as "*の*", were not marked. Function words were not included in t-tags.

2.2.2 Symptom and Diagnosis

Symptom and Diagnosis related expressions were marked with **<c> </c>** (c-tag) as follows.

Compound noun: each noun compound word was marked as a whole.

Verbal phrase: verbal phrases were not marked.

Body part and name of medical examination: the expressions of body parts, the disease names included in the medical examinations, and/or the names of the pathogenic bacteria of the examinations were not marked.

Disease identification test: if and only if the existence of the certain virus represents a single particular disease, the virus was marked.

ICD contained examination: despite the cases of medical examination, they were marked, if and only if the ICD codes were available.

Non-disease based phrase: the noun phrases that contained disease names as the parts of the medical tests and/or the surgeries were not marked.

General description: when the expression describes '*general'* information about the disease and/or the name of the clinic, it was not marked.

Non-alphabetical character: non-alphabetical characters and non-numerical characters (e.g. "↓") were marked with the previous noun phrases, if and only if the marks represented the conditions and were attached to the previous noun phrases to form the names of the disease.

Exemplifications: the expressions that describeed the degrees or tendencies of disease conditions, such as "*Teido 程度*," "*Hodo ほど*" and "*Keikou 傾向*," were marked. The expressions that described instabilities and variables, such as "*Tou 等*" and "*Nado など*," were not marked.

Modality related word: the words and phrases that suggested modalities of the symptoms (e.g. *positive 陽性*, *negative 陰性, prevention 予防, deterioration 悪化, emergence 出現, decline, depression 低下, enlarged 拡大, elevation 上昇, normal 正常, enlarge 拡張, diminution 縮小, progression 憎悪, change 変化,*

decrease 減少, recurrence 再発, continuum 継続, anamnestic 既往, ~able 可能) were not included in c-tags. Except, some modality-related words were included, if and only if they were connected to compose the standard disease names.

2.2.3 Modality Attribute

Patients' symptoms and diagnosis included 4 types of modalities are shown in Table 3.

In cases of the condition change for the better, they were marked with "negation." When diseases required two or more modalities, they were marked by separating each modality with commas (,).

2.2.4 ICD Attribute

Every symptom and diagnosis has its ICD-10 code. ICD-10 (The International Statistical Classification of Diseases and Related Health Problems 10th Revision) is a coding by WHO, that classifies diseases and signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases.

An ICD code is normally consisted of a single alphabet and 3 numerical digits. When the codes only had 2 or 1 digit(s), and/or when it was impossible to specify the codes from the contextual information, the last 1, 2 or 3 digits were supplemented with 1, 2 or 3 underbar(s) (_) to align ICD codes to be in 1-alphabet and 3-digits-forms.

3. METHODS

3.1 Task settings

In the NTCIR-11 MedNLP-2 task, we organized 3 types of tasks mentioned in Section 1. Task 1 and Task 2 required the following four steps.

Step 1: Corpus distribution: The sample set and the annotation guideline were sent to the participant groups for development.

Step 2: Task 1 submission: After two-month development period, the test set was sent to each participant group. Then the participant groups submitted their annotated results within a week. Multiple results with up to three systems were allowed to be submitted.

Step 3: Gold standard data of Task 1 distribution: After Task 1 submission, the gold standard data was sent to each participant group for Task 2.

Step 4: Task 2 submission: Task 2 participants annotated ICD-10 code to both their annotated data submitted in Task 1 (**Task 2 only track**) and gold standard data of Task 1 (**Task 2 only Track**). When the group participated Task 2 but not Task 1, the group would use the gold standard data of Task 1 provided after Task 1 deadline.

3.2 Evaluation metrics in Task 1

Performance of Task 1 (complaint and diagnosis task) was assessed using both the F-measure (β=1) [7], precision, recall, and accuracy. Precision is the percentage of correct named entities found by a participant's system. Recall is the percentage of named entities present in the corpus that were found by the system. A named entity is regarded as correct only if it was an exact match of the corresponding entity in the data file. The evaluation method is the same as that of the CoNLL-2000 shared task. A Perl script used for evaluation was obtained from the CoNLL-2000 website.

We adopted evaluation of two types, NER (only) and NER+modality (total). NER was complaint and diagnosis or not, that was, only named entity. Modality was including four types of modalities (positive, family, negation and suspicion).

3.3 Evaluation metrics in Task 2

ICD-10 code's structure is: a code is consisted of one alphabet <A-Z> and 1 to 3 digit Arabic numerals <1-9>. The first alphabet indicates its disease category. For example, <I> indicates "*Diseases of the circulatory system*".

We evaluated the added ICD codes (should be exact matches) with the following two levels: (1) of phrase, and (2) of document. Phrase level evaluation required both the NE unit and its ICD code. Document-level evaluation was the briefest evaluation, which judged the set of ICD code for each record.

The formula was as follows:

- Phrase level = (# of correct ICDs) / (# of ICDs)
- Document level = (# of correct ICDs) / (# of ICD types in documents)

4. RESULTS

4.1 Participating systems in Task 1

In all, 24 systems (of 10 groups) participated in Task 1 (extraction task). Modality attribute was added to tags by 23 systems. Nineteen systems (of 8 groups) submitted time expression tag. Table 4 shows that most participated systems employed the conditional Random fields (CRF), which is one of the most popular supervised machine learning techniques. Group F was the baseline system using CRF and the training corpus distributed by the MedNLP-2 organizers. In contrast to complaint and diagnosis term extraction, the methods used for detecting modalities varied (Table 5). Group A, C, F, G, and I used rule or regular expression pattern with some clue words that were manually build.

All but one groups used CRF for extracting complaint and diagnosis. In MedNLP-1, more diverse methods were employed, but all top 3 groups used CRF. This outcome could have affected the MedNLP-2 participants for their system selection procedure. In that case, it may be said that the knowledge from the previous challenge has been shared. However, in another aspect, we lost the diversity of the means, and the participants would have few choices in methods in Task 1. Since the result of Task 1 directly affected the result of Task 2, it seemed that they tried to achieve high performance as much as possible in Task 1, which eventually encouraged the participants to choose CRF as their means.

Several groups used unlabeled corpus to enhance distributed labeled corpus (the training corpus). Creating labeled medical and clinical corpus is very expensive, and therefore, such attempts of examining the potential of utilizing unlabeled corpus are important to make further development.

Table 4. Methods and language resources used in Task 1.

Group	Methods	Tools	Language resources
A	CRF		MEDIS Hyojun Byomei Master MEDIS Shintai Shoken Master
B	RNN Brown clustering	word2vec	MEDIS Hyojun Byomei Master
C	CRF Brown clustering		Wikipedia
D	CRF		
E	CRF		
F (baseline)	CRF		
G	CRF		
H	CRF		MEDIS Hyojun Byomei Master LSD, T-Jisyo, MeDRA/J, GSK2012–D Past state examinations of Medical Doctors
I	CRF		MEDIS Hyojun Byomei Master MeDRA, Byomei diagnosis list, MeSH terms, SNOMED CT
J	CRF Rule		MEDIS Hyojun Byomei Master ComeJisyoV5

Table 5. Methods and language resources to extract modalities in Task 1.

Group	Methods	Tools	Language resources
A	Regular expression pattern		Suffixes and prefixes extracted from MEDIS masters
B	RNN		
C	Rule		
D	CRF		
E	CRF		
F (baseline)	CRF		
G	Rule		Cue words manually built
H	SVM K-means		Japanese Web N-gram corpus (Google)
I	Rule		Cue words manually built
J	CRF		

Performances of extracting modality attribute task showed higher than the previous task, except for extracting '*negations*'. The results varied more than the extraction of "complaints and diagnosis" and "time expression". Performances in extracting modality attribute task were not well correlated with those in complaint and diagnosis. Differences in extracting performances for '*suspicion*' and '*family*' were larger than the others. Their increase from MedNLP-1 was also larger. To achieve high accuracy in extracting modality attributes, using rules and/or regular expression patterns with some clue words seemed to be necessary. The system for extracting '*negation*' and '*suspicion*' still expect some improvements. The performance in time expression extraction task decreased compared to MedNLP-1 for both the best and the average scores.

The extra language resources used for this task are medical dictionaries and text corpora. The popular dictionary resources used for this task were '*MEDIS Hyojun Byomei Master*'. On the other hand, various text corpora were used; e.g. Group C and G used Wikipedia for word clustering (brown clustering), Group H used the '*GSK Dummy Electronic Health Record Text Data (GSK2012-D)'* and past State Examinations of Medical Doctors on Web for their self-annotation method.

4.2 Performances in Task 1

For complaint and diagnosis extraction, the best system achieved the score of 83.95 in F-measure, and the average of all participating group scores was 78.39. Group A, C, and D showed good performances (Figure 1).

On the basis of modality attributes, the best and the average scores in F-measure were 78.10 and 70.60 for '*positive*' (Figure 2(a)), 76.77 and 68.68 for '*negation'* (Figure 2(b)), 60.61 and 44.38 for '*suspicion'* (Figure 2(c)), and 89.74 and 73.98 for '*family*' (Figure 2(d)). Good performance groups were A and I for *'positive'*, I, C, and A for "negation", G, C and E for "suspicion", and C and B for *'family'*.

For date and time related expression extraction, the best and the average scores in F-measure were 87.35 and 81.30. Group C, G, and D showed good performances (Figure 3).

Table 6. Methods and language resources in Task 2.

Group	Methods	Tools	Language resources
B	RNN	word2vec	MEDIS Hyojun Byomei Master ICD-10 English dictionary
C	SVN Brown clustering	word2vec	Wikipedia
D	Distance in tree hierarchy of ICD code's main and sub categories		MEDIS Hyojun Byomei Master
E	Full-test search Translation	Lucene Google translate	MEDIS Hyojun Byomei Master ICD-10 English dictionary
F (baseline)	Pattern match		MEDIS Hyojun Byomei Master
G	Pattern match Brown clustering		
H	Logistic regression		MEDIS Hyojun Byomei Master LSD, T-Jisyo, MeDRA/J
J	Rule		MEDIS Hyojun Byomei Master
K	Full-text search, Exact match Partial match、Feature-based match	Apache Solr	

4.3 Participating systems in Task 2

Most groups used pattern-matching algorithms for Task 2 (Table 6), but their methods varied more in Task 2 compared to Task 1, e.g. Group F (baseline system) and Group J used Levenshtein distance (Edit distance), Group E and K used Full text search system, Lucene and Solr. Team C and H used supervised machine learning methods, SVM and Logistic Regression.

4.4 Performances in Task 2

The methods the participants used for task 2 varied, which brought much divergence to their performance (Figure 4 and Figure 5). There were two tracks, (1) total track, which was the combination result of Task 1 and ICD coding, and (2) only coding track.

In total track (Task 1 & Task 2 participants) on the fine-grained evaluation (NE Level), the average was 0.532 (min: 0.251 - max: 0.676). On the rough evaluation (Document Level), the average was 0.614 (min: 0.297 - max: 0.771) (Figure 4).

In the ICD only track (Task 2 Only participants), on the fine-grained evaluation (NE Level), the average was 0.607 (min: 0.292 - max: 0.791). In the rough result (Document Level), the average was 0.644 (min: 0.317 - max: 0.823) (Figure 5).

In both tracks, the Group H (system H1 and H2) the highest performance was achieved.

4.5 Participating systems in Task 3

Task 3 was a task suggested by participants as practical or creative ideas other than Task 1 and Task 2. In MedNLP-2, two groups submitted their original tasks.

Group L examined the ratio of coverage on Medical Dictionary 2014 for ATOK IME 2014 which is a dictionary included in ATOK IME released by JustSystem.

Group F proposed a glossary of medical terms for patients. They extracted ICD-10 codes (appeared twice or more) from NTCIR corpus and investigated the words or expressions included in them. As a result, they got 316 terms in all and selected 98 terms as direction words from them to add a gloss to each term. In addition to the terms, they indexed the remaining 218 terms as related words of 98 direction words.

5. DISCUSSION

5.1 Overall performance in Task 1

Over all performances in extracting complaint and diagnosis increased from MedNLP-1 (though, the corpus and the annotation policy were not completely the same). Differences between high and low rank groups narrowed from MedNLP-1, and the number of groups who achieved higher score than the baseline system increased.

In MedNLP-2, the size of the training corpus has been three times larger than that of for MedNLP-1, and we had more writers for the corpus used in MedNLP-2. The size of the corpus had been increased, but the task was still difficult, since the test-set corpus was diverse and contained the terms that appeared neither in the training nor in the dry-run corpus. Having more writers would have made the corpus more diverse, and that would have been one reason why the degree of difficulty of MedNLP-2 would have been slightly increased compared to MedNLP-1.

In extracting “complaint and diagnosis” and “time expressions” tasks, the differences in performances of top groups were not so large. That suggested that the performance seemed to reach the plateau. However, these performances were still not satisfactory enough to utilize as an actual applications at hospitals. More novel approaches, abundant date, error analyses, and much more efforts for this task are required to meet the needs.

5.2 Instance level analysis of Task 1

Instance level analysis of Task 1 could have been a clue to improve systems by finding which named entity was difficult to extract. We calculated such degrees of difficulty for each named entity by counting how many systems failed to extract. We also checked frequencies of the named entities in the training corpus.

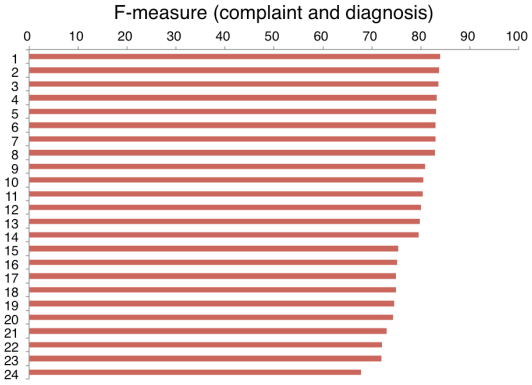

Figure 1. Performances in complaint and diagnosis extraction.

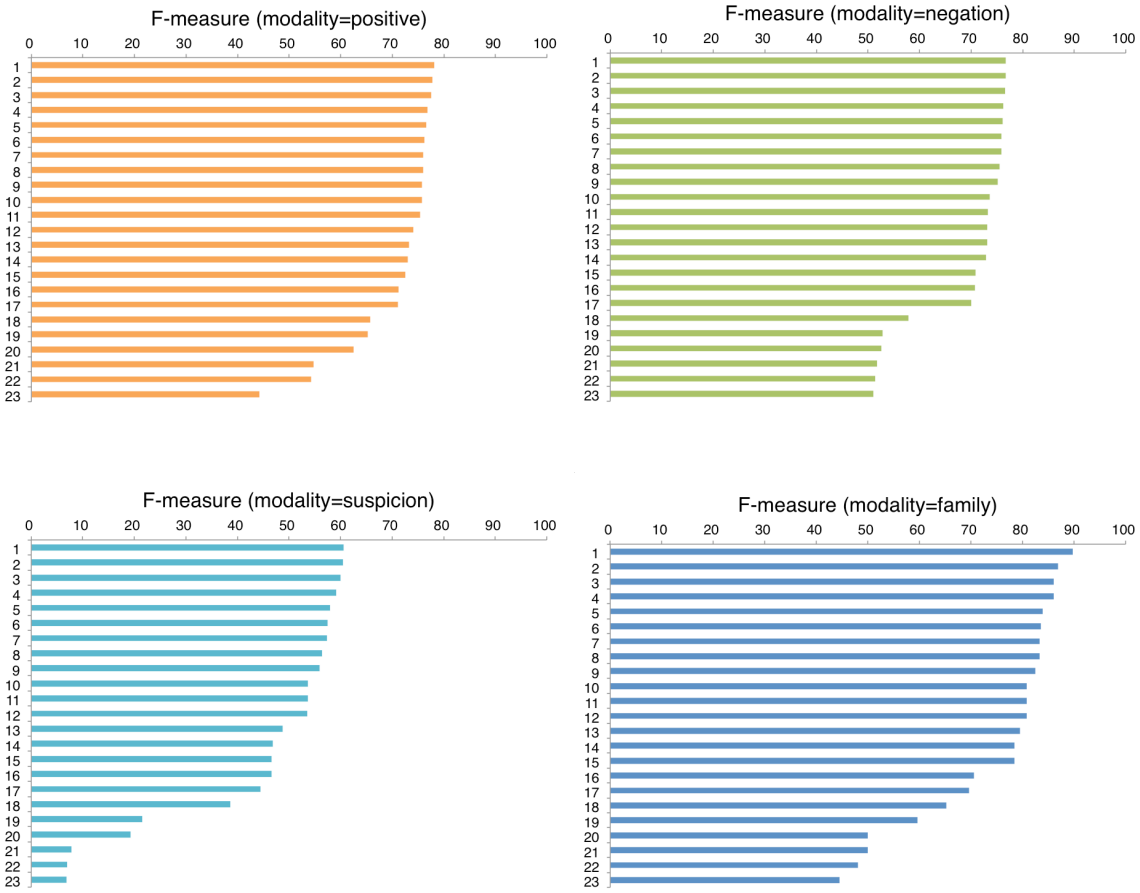

Figure 2. Performances in modality extraction: (a) positive, (b) negation, (c) suspicion, and (d) family.

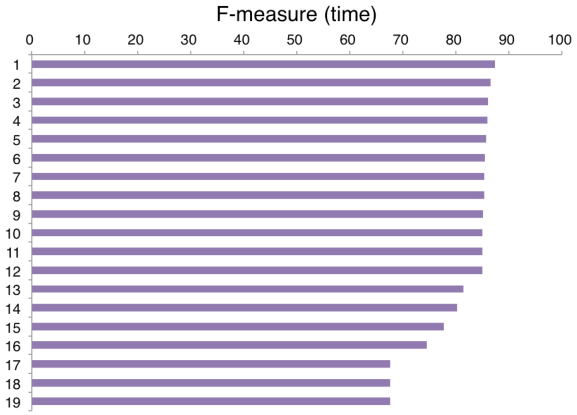

Figure 3. Performances in time expression extraction.

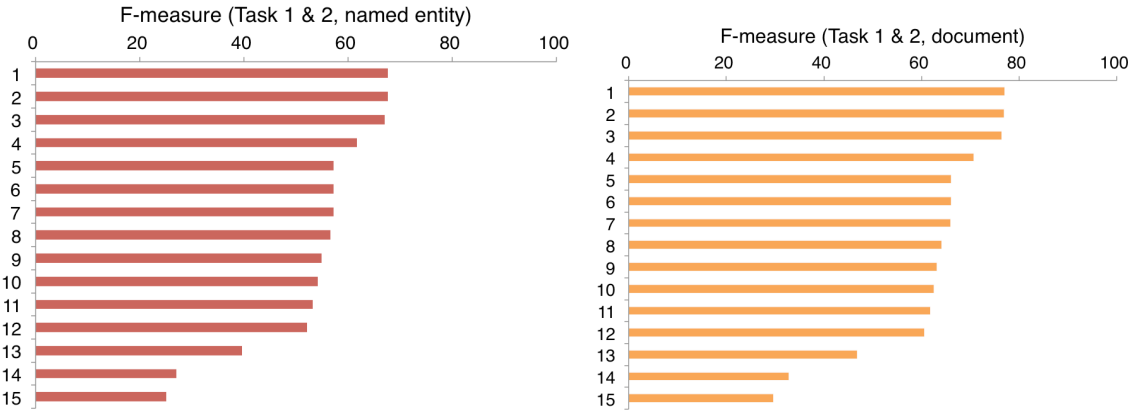

Figure 4. Performances in complaint and diagnosis normalization (ICD-10 coding) (Task 1 & 2): (a) evaluated on a named entity basis, (b) evaluated on a document basis.

Figure 5. Performances in complaint and diagnosis normalization (ICD-10 coding) (Task 2 only): (a) evaluated on a named entity basis, and (b) evaluated on a document basis.

The most straightforward observation was the correlation between the difficulty of named entities and its corresponding frequency in the training corpus. There were 138 occurrences of named entities, to which no team could give correct answers. These 138 occurrences may be regarded as the most difficult entities to extract. Among these entities, 13 entities had not appeared in the training corpus. These 13 entities were; "*洞調律*", "*呼吸苦*", "*脾*", and "*痰*". In the training corpus, "*洞調律*" mostly appeared as "*正常洞調律*" without any annotation marked. Morphological analyzers would have divided "*正常洞調律*" into "*正常*" and "*洞調律*", then the O tags of BIO would have be assigned, which might have lead to the wrong training result. Registering "*正常洞調律*" to the dictionary may have solved this sort of problem. In contrast, "*呼吸苦*" always appeared with annotations marking this very entity span in the training corpus. Morphological analyzers might have divided this word into "*呼吸*" and "*苦*", and BIO tag distribution went biased. "*脾*" and "*痰*" were too generic that appeared within larger morphemes. Overall, morpheme division should have been one of the critical issues, especially in the CRF based methods.

On the other hand, many NEs that did not appear in the training corpus were correctly detected. These NEs appeared together with similar patterns of neighboring morphemes. This result showed benefits of CRF based methods that may be able to learn this sort of contexts.

We have observed diversity of which system failed on which NE. For example, NEs, where a single system correctly answered, could have been simply covered by a single system if this was a problem of CRF tuning. However, many different systems were observed and gave correct answers, while other systems failed. It may have reflected the differences of their dictionaries or the rules, if any were employed.

5.3 Overall performance in Task 2

The ICD coding task was newly introduced to MedNLP-2. The ICD code covers thousands of categories that represented most diseases known today, and thus, was a new challenge for NLP. The participants used varieties of approaches to challenge this task.

Their approaches were classified into two types: (1) supervised approaches and (2) non-supervised approaches. In supervised approaches, SVM and Logistic regression, which are popular methods in classification, were employed. In unsupervised approaches, string similarity measures (such as edit distance) were utilized. Several groups employed heuristic based similarity using soft match manner, partial match manner, prefix-suffix, and so on.

The option of extra renounces was diverse. Several groups utilized the combination of many resources, such as MEDIS Hyojun Byomei Master, LSD, T-terminology Dictionary, MedDRA/J, and the other groups did not utilize any extra resources.

These varieties in approaches and resources have generated much divergence in performance. The lowest performance was 0.292, and the best performance was 0.791 in Task 2 Only track.

Among all evaluations, the Group H achieved the highest performance. Note that the Group H did not show the best result in Task 1, but the group achieved the highest score in total. The method employed in Group H was based on supervised learning manner, which is commonly used for classification tasks like Task 2. The originality of their system was their use of the extra resources, such as MEDIS, LSD, TDIC and MedDRA. Especially, LSD, TDIC and MedDRA were the unique resources used by the Group H, which suggested that these resources might have strongly contributed to the Task 2 performance.

Group B employed a relatively new technique, word2vec. Although the performance of Group B was not so high, it gave the new insight and challenge for the further investigation.

6. CONCLUSION

This paper describes an overview of the NTCIR-11 MedNLP-2 task. The MedNLP-2 task was our second attempt to analyze medical documents written in Japanese by using fair evaluation techniques. The total of 12 different groups participated in MedNLP-2, which included three subtasks. In extraction of medical terms task, which was subsequent to MedNLP-1, the methods used by the participants were similar to each other, and thus, their result scores came up in a narrow range. New technical breakthroughs are needed to be explored for further performance. In normalization of medical terms task, various methods were applied, and the range of result scores was wide. That implies that the measures to challenge this task were still at the primitive level. However, the top systems achieved high scores and showed much potential. We will continue producing the community of developers and stakeholders by constructing new tasks for them to participate. In addition, we pursue developing more practical tools and the components that are to be used in medical natural language processing.

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

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