Overview of the NTCIR-10 MedNLP Task

Mizuki Morita
The University of Tokyo
morita@cks.u-tokyo.ac.jp

Yoshinobu Kano
JST PRESTO
kano@nii.ac.jp

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

Mai Miyabe
The University of Tokyo
mai.miyabe@gmail.com

Eiji Aramaki
The University of Tokyo
eiji.aramaki@gmail.com

ABSTRACT
Recently, medical records are increasingly written on electronic media instead of on paper, thereby increasing the importance of information processing in medical fields. We have organized an NTCIR-10 pilot task for medical records. Our pilot task, MedNLP, comprises three tasks: (1) *de-identification*, (2) *complaint and diagnosis*, and (3) *free*. These tasks represent elemental technologies 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). This report presents results of these systems, with discussion clarifying the issues to be resolved in medical NLP fields.

Keywords
medical records, electronic health records (EHR), de-identification, named entity recognition (NER), shared task and evaluation

1. INTRODUCTION
Medical records are increasingly written on electronic media instead of on paper, which has radically increased the importance of information processing techniques in medical fields. Nevertheless, the state of usage of information and communication technologies (ICT) in medical fields is said to 10 years behind that in other fields. By processing large amounts of medical records and obtaining knowledge from them, great potential exist in assisting more precise and timely treatments. Such assistance can save lives and provide better quality of life.

Our goal is the promotion and support of implementation of practical tools and systems in the medical industry, which can support medical decisions and treatment by physicians and medical staff. A short-term objective of this pilot task is to evaluate basic techniques of information extraction in medical fields, but the long-term objective is to offer a forum for achieving the goal with a community-based approach. We aim to gather people who are interested in this issue. Then we intend to facilitate their communication and discussion to clarify issues to be solved, while defining the necessary elemental technologies.

Numerous community-based attempts called 'shared task' (or contest, competition, challenge evaluation, critical assessment) encourage research in information retrieval. Among these shared tasks, the best-known shared task specifically related to medical fields is Informatics for Integrating Biology and the Bedside (i2b2)[1], started in 2006 by the National Institutes of Health (NIH). The Text Retrieval Conference (TREC)[2], which deals with diverse issues, also launched the Medical Records Track in TREC 2011. Both shared tasks are targeting at English text. Because medical records are written in native languages in most countries, information retrieval techniques must be developed for individual languages.

The NTCIR-10 MedNLP Task is a shared task that evaluates technologies to retrieve important information from medical reports written in Japanese. This is the first attempt of shared tasks targeting at medical documents in Japanese. Although the tasks in this attempt are apparently basic, they are related to elemental and important technologies for developing computational systems that support widely various medical applications.

2. TASK DESCRIPTION
2.1 Data Preparation
We have created medical history summary reports, written in Japanese by physicians, of patients with putative or diagnosed diseases. Medical records contain extremely sensitive personal information about patients and others such as patients' families, friends, and colleagues. Therefore, we asked physicians to write

Table I. Tags in medical records

Tag	Description*
(a) Personal information tag	
<a>	age (56)
<p>	person's name (0)
<x>	sex (4)
<t>	time (355)
<h>	hospital name (75)
<l>	location (2)
(b) Medical information tag	
<c>	complaint and diagnosis (1,922)

* Parentheses show numbers of tags in 2,244 sentences

Table II. Modalities within the <c> tag

Modality	Description*
positive†	positive finding in the patient (1,314)
family	positive finding in the person's family (32)
negation	negative finding (504)
suspicion	suspicious finding (72)

* Parentheses show numbers of modalities in 1,922 sentences

† This modality was omitted from the *sample set*.

[1] http://www.i2b2.org/

[2] http://trec.nist.gov/

Table III. Short description of groups participating in the NTCIR-10 MedNLP Task

Group ID	Methods	Resources
NTTD	Word match, Semi-supervised learning	MEDIS Standard Masters (disease names)
LSDP	N/A	N/A
KobeU	Structured perceptron	
ulab	Online learning	Japanese newspaper (disease names)
msiknowledge	CRF, Language Model	
UT-FX	CRF	MedDRA/J, MEDIS Standard Masters, Original corpus
HCRL	CRF, Word match	Japanese Wikipedia (disease names)
niph	Word match	Original dictionary
oka1	CRF, Word match	
NECLA	CRF	UMLS, LSD
cks01	CRF	MEDIS Standard Masters
SinicaNLP	Word match, Machine translation	Original dictionaries (in Chinese)

down fictional medical reports of imaginary patients. Each medical report typically contains a chief complaint, a past medical history, diagnosis, treatments, clinical course, and outcome. We offered 50 collected medical reports for this task, which include 3,365 sentences in all: about 40,000 words.

We annotated both personal and medical information in these medical reports (Table I) according to our annotation guideline. Personal information includes age, person's name, sex, time, hospital name, and location, which are tagged respectively as <a>, <p>, <x>, <t>, <h>, and <l>. The medical information is the complaint and diagnosis, which are tagged with <c>. As shown in Table II, we defined a modality attribute for <c>, which has attribute-values of four kinds. They are “positive”, “suspicious”, “negative”, and “family.” The value "positive" expresses a doctor’s confidence for a phrase marked by <c>. The value is set as a default value in the corpus so a modality attribute and its value are omitted. The value "suspicious" expresses a doctor’s uncertainty about a phrase marked by <c>. For example, the attribute-value of "breast cancer" is "positive” in "The patient had <c>breast cancer</c>." However, the attribute-value of "breast cancer" is "suspicious" in “The patient is suspected of <c modality=”suspicious”>breast cancer</c>.” Because a doctor is not confident of the diagnosis. The attribute-value "negation" expresses a phrase marked by <c>, which does not hold. For example, the attribute-value of "breast cancer" is "negative" in "The patient has no <c modality=”negative”>breast cancer</c>." The value "family" expresses the patient's family medical history.

工場に勤めている<a>64歳の<x>男性</x>。<t>2025年月8月2日（来院5日前）頃から</t><c>腹痛</c>が生じるとともに，<c>食欲不振</c>，<c>嘔気</c>・<c>嘔吐出現</c>した。体幹は温かいが，末梢は<c>湿潤冷汗</c>で<c>ショック状態</c>。明らかな<c modality="negation">運動麻痺</c>はみられず。<t>翌日</t>，<c>意識障害出現</c>し，<c>腎機能障害</c>の増悪を認めて徐々に<c>尿量低下</c>し，<t>8月9日18時10分</t>に<c>心肺停止</c>。<t>8月9日21時44分</t><c>死亡確認</c>。

Figure 1. Example of annotated sentences.

For example, the attribute-value of "breast cancer" is "family" in "The patient's mother had <c modality=”family”>breast cancer</c>". An example of an annotated medical report with these tags is presented in Figure 1.

The whole annotated corpus was split after shuffling sentences randomly into a *sample set* (including 2,244 sentences) for development and a *test set* (including 1,121 sentences) for evaluation of participating systems.

2.2 Subtasks and Timeline

In the NTCIR-10 MedNLP Task, we have organized the following tasks of three types:

1. *De-identification tas*k: this task adds personal information tags to the *test set.*
2. *Complaint and diagnosis task*: this task adds the patient status information tag to the *test set.*
3. *Free task*: tasks suggested by participants as practical or creative ideas other than the tasks described above.

Both Task 1 (*de-identification task*) and Task 2 (*complaint and diagnosis task*) can be regarded as a variation named entity recognition (NER). However, these tasks include inherent difficulties compared to other standard NER tasks because medical records are written mostly in an unstructured and ungrammatical manner.

Before the registration deadline of the MedNLP Task, an example (Fig. 1) was publicly provided on the shared task website together with the call for participation. After the registration closed, the *sample set* and the annotation guideline were sent to the participant groups for development. After a two-month development period, the *test set* was sent to the participant groups. Then the participant groups were required to submit their annotated results within a week. Each group was allowed to submit multiple results with up to three systems.

2.3 Evaluation Metrics

Performance of Task 1 (*de-identification task*) and Task 2 (*complaint and diagnosis task*) is assessed using the *F*-measure.

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{(\beta^2 * precision + recall)}$$

Table IV. (a) Overall results and (b) detailed results for each privacy type in Task 1 (*De-identification task*)

(a)

	P	R	F	A
C3	89.59	91.67	90.62	99.58
B3	91.67	86.57	89.05	99.54
B1	90.05	87.96	88.99	99.49
B2	90.82	87.04	88.89	99.52
C1	92.42	84.72	88.41	99.49
A1	91.50	84.72	87.98	99.47
C2	91.50	84.72	87.98	99.46
A2	90.15	84.72	87.35	99.41
D1	86.10	74.54	79.90	99.36
G1	82.09	76.39	79.14	99.38
D3	85.87	73.15	79.00	99.35
D2	80.81	74.07	77.29	99.24
H2	76.17	75.46	75.81	99.28
H1	75.81	75.46	75.64	99.27
H3	74.88	74.54	74.71	99.26

(b)

	<a> age			<x> sex			<t> time			<h> hospital name		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
C3	90.32	87.5	88.89	100	100	100	87.16	91.49	89.27	97.30	94.74	96.00
B3	90.00	84.38	87.10	100	50.00	66.67	91.30	89.36	90.32	97.06	86.84	91.67
B1	93.33	87.5	90.32	100	100	100	90.65	89.36	90.00	89.47	89.47	89.47
B2	90.00	84.38	87.10	100	100	100	91.24	88.65	89.93	91.89	89.47	90.67
C1	96.67	90.62	93.55	100	50.00	66.67	91.18	87.94	89.53	93.55	76.32	84.06
A1	92.86	81.25	86.67	100	50.00	66.67	91.04	86.52	88.73	91.89	89.47	90.67
C2	96.67	90.62	93.55	100	50.00	66.67	89.13	87.23	88.17	96.77	78.95	86.96
A2	92.86	81.25	86.67	100	50.00	66.67	89.05	86.52	87.77	91.89	89.47	90.67
D1	92.31	75.00	82.76	100	50.00	66.67	82.84	78.72	80.73	96.15	65.79	78.12
G1	80.65	78.12	79.37	100	50.00	66.67	84.56	81.56	83.03	72.73	63.16	67.61
D3	88.89	75.00	81.36	100	50.00	66.67	83.08	76.60	79.70	96.15	65.79	78.12
D2	92.31	75.00	82.76	100	50.00	66.67	75.86	78.01	76.92	96.15	65.79	78.12
H2	83.87	81.25	82.54	100	100	100	73.79	75.89	74.83	77.78	73.68	75.68
H1	80.65	78.12	79.37	100	100	100	75.86	78.01	76.92	70.27	68.42	69.33
H3	83.87	81.25	82.54	100	100	100	73.79	75.89	74.83	70.27	68.42	69.33

P, precision; *R*, recall; *F*, *F*-measure (β=1); and *A*, accuracy. *P*, *R* and *F* were calculated at the phrase level. A was calculated in the word level (the agreement ratio of B-*, I-* and O).

In that equation, β=1 [1]. 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 are found by the system. A named entity is regarded as correct only if it is 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[3].

2.4 Participating Systems

In all, 15 systems (from 6 groups) have participated in Task 1 (*de-identification task*), 22 systems (from 11 groups) in Task 2 (*complaint and diagnosis task*), and 1 system (from 1 group) in Task 3 (*free task*). For Tasks 1 and 2, most participating systems have used machine learning techniques and language resources other than the available training data, such as medical dictionaries and extra annotated or un-annotated data.

Free task is a new challenge for this pilot task. Because various participants such as medical doctors, engineers, and computer scientists have a great variety of final goals, we have organized a *free task* as an unrestricted track. Consequently, this track has one participant: LSDP. This team, which has been developing an English–Japanese thesaurus of medical terms named Life Science Dictionary (LSD) for 20 years, investigated the coverage of their original dictionary using our corpus, and reported matching results.

2.5 Techniques and Methods

The most frequently applied technique to both the *de-identification task* and the *complaint and diagnosis task* is the Conditional Random Fields (CRF) Model [2]. Six groups used

[3] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

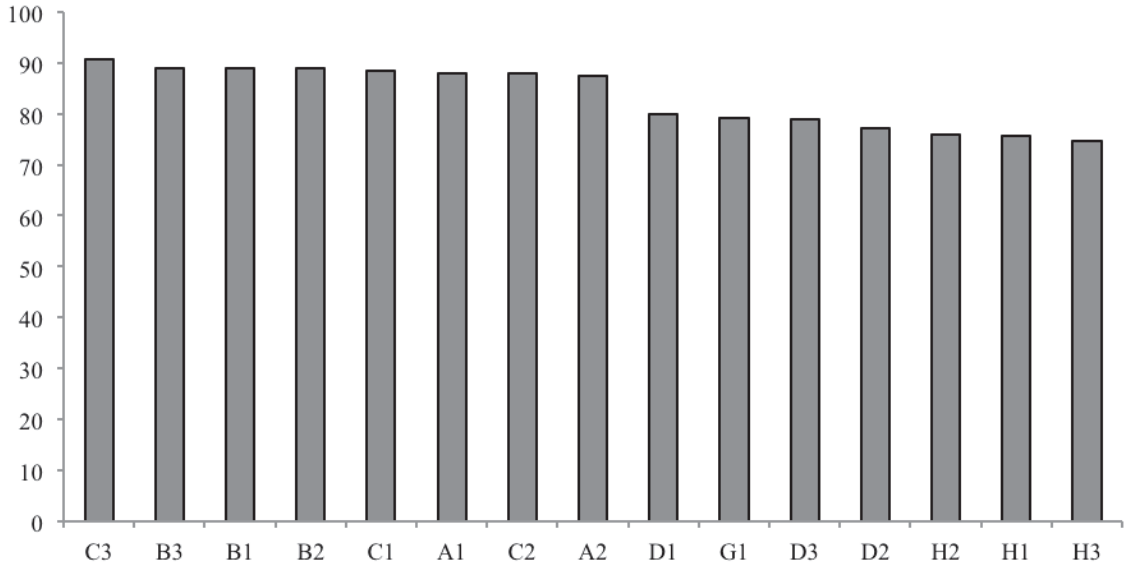

Figure 2. Performance of all systems in *F*-measure in Task 1 (*De-identification task*).

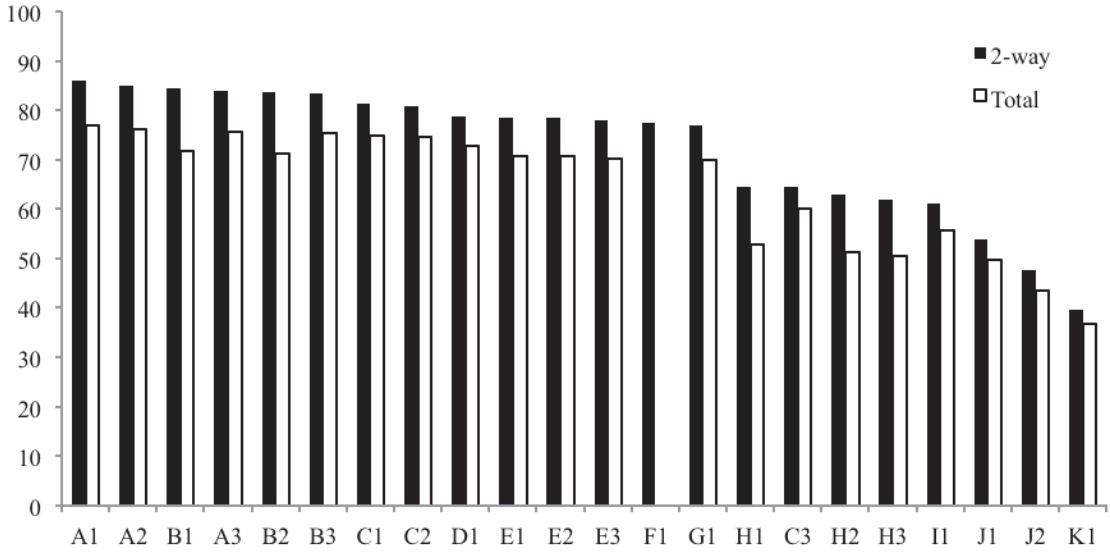

Figure 3. Performance of all systems in *F*-measure in Task 2 (*Complaint and diagnosis task*).

this statistical learning method. KobeU has applied the other technique (structured perceptron).

CRF is apparently useful for the *complaint and diagnosis task*: The top three systems for the *complaint and diagnosis task* have used this technique. However, the best system for the *de-identification task* has used a rule-based technique, whereas the other top systems have used CRF.

Three groups applied word-matching techniques. SinicaNLP used Chinese resources by translating the *test set* from Japanese into Chinese.

2.6 Language Resources

Table III presents a short description of each group's approach. Seven groups used extra dictionaries, which are classified into two types. One is publicly available dictionaries such as MEDIS Standard Masters[4], Unified Medical Language System (UMLS)[5], and Life Science Dictionary (LSD)[6]. Another is in-house dictionaries. Clinical term dictionaries developed manually by a

[4] http://www.medis.or.jp/4_hyojyun/medis-master/

[5] http://www.nlm.nih.gov/research/umls/

[6] http://lsd.pharm.kyoto-u.ac.jp/

Table V. (a) Overall results and (b) detailed results for each modality type in Task 2 (*Complaint and diagnosis task*)

(a)

	2-way				Total			
	P	R	F	A	P	R	F	A
A1	90.73	81.60	85.93	96.59	81.23	73.05	76.92	95.39
A2	88.34	82.03	85.07	96.46	79.02	73.38	76.09	95.26
B1	89.68	79.98	84.55	96.43	75.97	67.75	71.62	94.74
A3	88.26	79.76	83.80	96.33	79.76	72.08	75.72	95.22
B2	89.01	78.90	83.65	96.24	75.70	67.10	71.14	94.55
B3	89.76	77.81	83.36	96.37	81.15	70.35	75.36	95.26
C1	88.55	75.32	81.40	96.06	81.42	69.26	74.85	95.15
C2	88.98	74.24	80.94	96.08	82.10	68.51	74.69	95.22
D1	87.37	71.86	78.86	95.91	82.29	65.37	72.86	94.87
E1	78.91	78.14	78.52	94.57	71.15	70.45	70.80	93.46
E2	79.44	77.38	78.40	94.56	71.56	69.70	70.61	93.45
E3	80.00	76.19	78.05	94.43	71.93	68.51	70.18	93.32
F1	86.52	70.13	77.47	95.74	-	-	-	-
G1	82.37	72.29	77.00	95.48	74.72	65.58	69.86	94.50
H1	66.32	62.88	64.56	93.72	54.34	51.52	52.89	91.82
C3	72.47	58.12	64.50	93.40	67.61	54.22	60.18	92.83
H2	64.86	60.93	62.83	93.41	53.0	49.78	51.34	91.56
H3	63.32	60.71	61.99	93.29	51.58	49.46	50.50	91.43
I1	58.67	63.74	61.10	93.50	53.49	58.12	55.71	92.49
J1	54.75	53.03	53.88	91.46	50.39	48.81	49.59	90.80
J2	51.84	44.16	47.69	91.09	47.40	40.37	43.60	90.47
K1	58.60	29.87	39.57	90.21	54.35	27.71	36.70	89.76

(b)

	c-*positive*			c-*family*			c-*negation*			c-*suspicion*		
	P	R	F	P	R	F	P	R	F	P	R	F
A1	81.91	76.80	79.27	84.62	50.00	62.86	80.91	72.06	76.23	50.00	20.00	28.57
A2	79.02	77.12	78.06	83.33	45.45	58.82	80.18	72.06	75.91	57.14	26.67	36.36
B1	80.80	68.00	73.85	65.22	68.18	66.67	75.23	67.61	71.22	35.85	63.33	45.78
A3	79.16	75.36	77.21	76.92	45.45	57.14	82.41	72.06	76.89	63.64	23.33	34.15
B2	80.61	67.20	73.30	71.43	68.18	69.77	74.32	66.80	70.36	36.36	66.67	47.06
B3	80.92	73.28	76.91	82.35	63.64	71.79	84.50	68.42	75.62	50.00	30.00	37.50
C1	78.84	71.52	75.00	100	68.18	81.08	89.47	68.83	77.80	57.14	26.67	36.36
C2	79.61	71.20	75.17	100	68.18	81.08	90.16	66.80	76.74	57.14	26.67	36.36
D1	79.66	67.04	72.81	100	54.55	70.59	90.16	66.80	76.74	61.54	26.67	37.21
E1	73.04	70.24	71.62	77.27	77.27	77.27	71.26	71.26	71.26	42.22	63.33	50.67
E2	73.51	69.28	71.33	77.27	77.27	77.27	72.02	70.85	71.43	41.30	63.33	50.00
E3	74.39	68.80	71.49	77.27	77.27	77.27	71.37	67.61	69.44	41.30	63.33	50.00
F1	-	-	-	-	-	-	-	-	-	-	-	-
G1	72.87	67.04	69.83	66.67	36.36	47.06	82.35	68.02	74.50	55.00	36.67	44.00
H1	57.39	54.08	55.68	42.11	36.36	39.02	51.20	51.82	51.51	11.11	6.670	8.33
C3	68.21	54.24	60.43	100	63.64	77.78	69.79	54.25	61.05	36.84	46.67	41.18
H2	56.03	52.00	53.94	40.00	45.45	42.55	50.62	49.80	50.20	10.00	6.67	8.00
H3	54.01	51.68	52.82	40.00	45.45	42.55	50.21	49.39	49.80	10.00	6.67	8.00
I1	48.86	61.76	54.56	70.83	77.27	73.91	73.05	49.39	58.94	52.17	40.00	45.28
J1	49.53	50.72	50.12	78.95	68.18	73.17	53.61	42.11	47.17	35.71	50.00	41.67
J2	47.20	43.20	45.11	78.95	68.18	73.17	47.13	29.96	36.63	35.90	46.67	40.58
K1	54.70	35.36	42.95	100	18.18	30.77	50.00	10.93	17.94	44.44	13.33	20.51

P, precision; R, recall; F, F-measure (β=1); A, accuracy. P, R, and F were calculated in the phrase level. A was calculated in the word level (the agreement ratio of B-*, I-* and O). The two-way result means a simple result (complain/diagnosis or not). The total result means more fine-grained result including modality (*positive*, *family*, *negation*, and *suspicion*).

physician were used by niph. Automatically compiled dictionaries have also been applied: HCRL collected terms from Japanese Wikipedia using clustering algorithm; ulab extracted disease names from newspaper; and NTTD applied a bootstrapping method (Espresso [3]) for extraction of disease names.

3. RESULTS OVERVIEW

3.1 Overall Results of Task 1: De-identification Task

Table IV(a) and Figure 2 present overview results in Task 1 (*de-identification task*). The top systems achieved high performance (around 90% of *F*-measure). Considering the small corpus size of this pilot task, this result demonstrates the fundamental feasibility of automatic de-identification.

Table IV(b) presents detailed results in Task 1 (*de-identification task*). Among the various privacy information types, the hospital name (<h>) was easy to detect. One reason is that the hospital names share typical expressions, such as "… *Hospital*", "… *Clinic*", and "… *Center*".

Most systems (from system C3 to A2 in Table IV(a) and Figure 2) show no statistically remarkable difference in detecting the time tag (<t>), which implies that they share the same difficulty left to solve, although the approaches and resources differ.

The time tag has the largest number of annotations. Therefore, the overall evaluation scores closely correlate with the time detection scores. However, age (<a>) detection results show a somewhat different tendency. Even in the top groups, the *F*-measures are 86–93, not correlating with the total scores, which implies that age detection has a different difficulty than others such as missing age-specific expressions.

3.2 Overall Results of Task 2: Complaint and Diagnosis Task

Table V(a) and Figure 3 present overview results in Task 2 (*complaint and diagnosis task*). The left-hand side of Table V(a) shows two-way results (evaluation results without considering modality attributes). The top three systems (A, B and C) achieved high performance (around 85% of *F*-measure), which indicates that complaints are also easy to detect. However, modality-classifying results (shown in the right of Table V(a)) show poor performance (under 77% of *F*-measure), which is almost 10 points lower than two-way results, indicating that modality classification is a difficult task for current NLP systems.

Table V(b) present detailed results for Task 2 (*complaint and diagnosis task*) for each modality type. Among the various modalities, negation detection shows higher performance. One reason is that negation is a popular phenomenon (the corpus contains many negation examples). The positive modality also showed higher performance, perhaps attributable to the greater number of annotations given in the training corpus.

The family modality includes unique systems. C1, C2, D1, and C3 obtained 100% precision; their *F*-measures are also the best ones. This result implies that targeting at higher precision will improve the entire system, perhaps by using a good dictionary.

The suspicion modality produced extremely low performance, requiring additional studies.

4. FUTURE WORKS

The participants' systems can be more useful when they become ready to use, which means not just achieving ease at operating the systems, but also compatibility with existing systems. For these purposes, we plan to incorporate some of the participants' systems into an Unstructured Information Management Architecture (UIMA) framework [4] and make them Kachako [5] compatible. Kachako is an integrated NLP platform that can run UIMA components, offering full automation features from installation to large scale processing. Once a MedNLP system becomes a Kachako component as free software, users can easily run a workflow using that MedNLP component in their own local environment, e.g. processing sensitive medical records without sending any information outside.

5. CONCLUSION

This paper described an overview of the NTCIR-10 MedNLP Pilot Task. The MedNLP Task is the first attempt to analyze medical documents in Japanese using fair evaluation techniques. MedNLP included three subtasks, in which a total of 38 systems from 12 different groups had participated.

Although the *sample set* would not contain sufficient texts for training of machine learning methods, several groups achieved higher scores than we had anticipated. We were pleased to see novel approaches pointing in the direction of future development.

We will continue our efforts with new and continuing tasks to produce a community of developers and stakeholders. We also intend to develop practical tools and their components for use in medical natural language processing (MedNLP).

6. REFERENCES

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

[2] J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282-289, 2001.

[3] P. Pantel and M. Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113-120, 2006.

[4] D. Ferrucci , A. Lally, D. Gruhl , E. Epstein, M. Schor, J. W. Murdock, , A. Frenkiel, E. W. Brown, T. Hampp et al. Towards an Interoperability Standard for Text and Multi-Modal Analytics. *IBM Research Report*, 2006.

[5] Y. Kano. Kachako: a Hybrid-Cloud Unstructured Information Platform for Full Automation of Service Composition, Scalable Deployment and Evaluation. In the *1st International Workshop on Analytics Services on the Cloud (ASC), the 10th International Conference on Services Oriented Computing (ICSOC 2012),* pages 72-84, 2012.