

Technical Report of Uni2014 in NTCIR-11 MedNLP-2 (Extraction and Normalization Task)

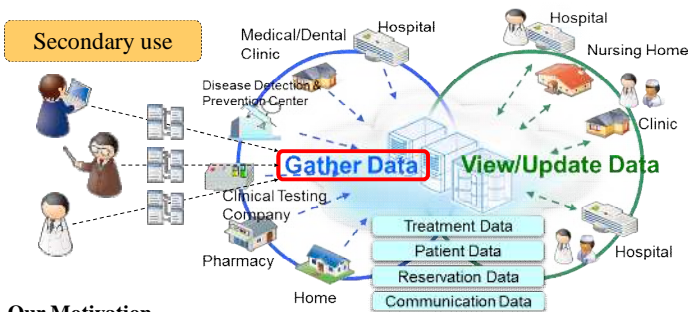
Kenta Fukuda

Nihon Unisys Technology R&D Center

Introduction

Our Approach in healthcare business

- Nihon Unisys is promoting information and communication technology (ICT) development based on business results of information platform construction related to medical care and health.



Our Motivation

- Our next step in business is to secondarily and beneficially utilize the data which gather from healthcare facilities. Our Motivation is to research and develop basic techniques to information extraction in medical fields, and review the availability to our medical information system solutions.

Methods

Dataset and Task Image

Input Data Sample

8月2日頃から腹痛が生じるとともに、嘔気・嘔吐出現

Output Data Sample

<t>8月2日頃</t>から <c icd="R104">腹痛</c>が生じるとともに、<c icd="R11_">嘔気</c>・<c icd="R11_">嘔吐</c>出現

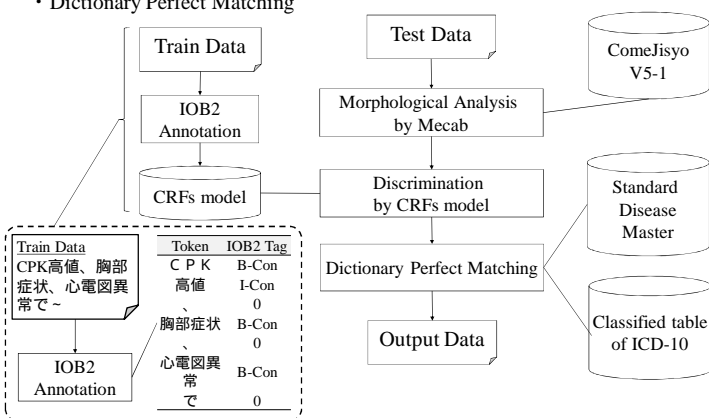
Task 1) Extract complaint and diagnosis, and it's modality attributes (modality="negation", "suspicion", "family")

Task 2) Give ICD-10 code on complaint and diagnosis

Method design in Task1 (Figure.1)

- Morphological Analysis by Mecab
- Discrimination by CRFs model
- Dictionary Perfect Matching

Figure.1



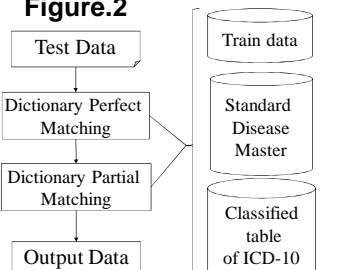
Method design in Task2 (Figure.2)

- Replacement when
- Perfect Matching
 - Partial Matching

Using three dictionary of

- Train data
- Standard Disease Master
- ICD-10

Figure.2



Results

Task1 (Table.1)

- We could extracted time and condition(only) related expressions in a high accuracy. On the other hand it was difficult to extract the modality types, especially "Suspicion".

Table.1

IOB2 Tag	Precision	Recall	F
Time	88.14	74.53	74.53
Condition(Only)	81.83	69.38	75.09
Condition(Positive)	70.17	56.34	62.50
Condition(Negative)	51.05	54.61	52.77
Condition(Suspicion)	45.45	12.20	19.23
Condition(Family)	66.67	53.85	59.57

Task2 (Table.2)

- The result of Task2 is shown in Table 2. We used "GoldStandard+ICD" as test data which was distributed by MedNLP-2 organizer as correct answer data of Task1.

Table.2

Data	Accuracy (%)
GoldStandard+ICD	69.40

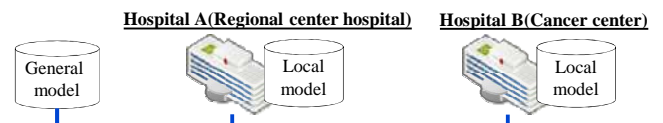
Discussions

Modality Extraction in Task1

- It was difficult to extract the modality types in a high accuracy. We think that adding more rules to complement our algorithm is needed.

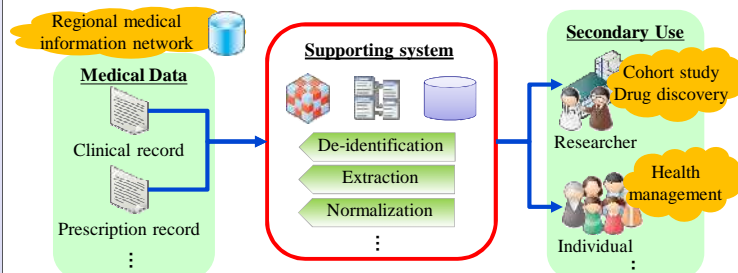
Consideration of Local recording rule in each hospitals

- It is important to consider the local rule of each hospitals and add these rules to the algorithm to use widely as a function of medical information system, because each hospital have each rule of recording document.



For beneficial Data Mining

- It is very important not only to develop these supporting system but also to consider a purpose to analysis data which are extracted from medical data, because the requirements of extracting data depend strongly on the purposes of data analysis. We think that each approaches are needed.



Bibliography

- Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T. 2014. Overview of the NTCIR-11 MedNLP-2 Task. In Proceedings of the 11th NTCIR Workshop Meeting on Evaluation of Information Access Technologies.
- Morita, M., Kano, Y., Ohkuma, T., Miyabe, M., and Aramaki, E. 2013. Overview of the NTCIR-10 MedNLP Task.
- Kudo, T., Yamamoto, K., Matsumoto, Y. 2004. Applying Conditional Random Fields to Japanese