財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# NTCIR-11 MedNLP

## Preliminary Report of III&CYUT for NTCIR-11 MedNLP-2

Liang-Pu Chen†, Hsiang-Lun Lin†, Yan-Shen Lai†, Ping-Che Yang†*
†IDEAS, Institute for Information Industry, Taiwan
*corresponding author
{eit, hsianglunlin, larslai, maciaclack} @iii.org.tw

**Abstract**

We construct a supervised learning system to participate MedNLP2 task in NTCIR-11 that find the keyword out correctly at right position and normalize to identify unique id in ICD10. In our system, We pick part-of-speech tagging (POS) as feature to train machine learning models based on Conditional Random Fields (CRF) for named entities extraction, then construct a hierarchical classifier to determine ICD code of the terms.

## Method

We have two process in our system. The first is training, and the next is analysis. The picture as below is the structure of training model. In training process, system will split our resource into word table and find their speech by POS system. The system will merge them to become one table and use our template to train CRF to become our training model. The next process is analysis. The picture as below is the structure of analysis model. In analysis process, the system will split the context into word table and analysis the speech by POS system. Use CRF system to get the keyword and reform to new context which have keyword tag. In process, we have two major steps, format conversion and symptom recognition.

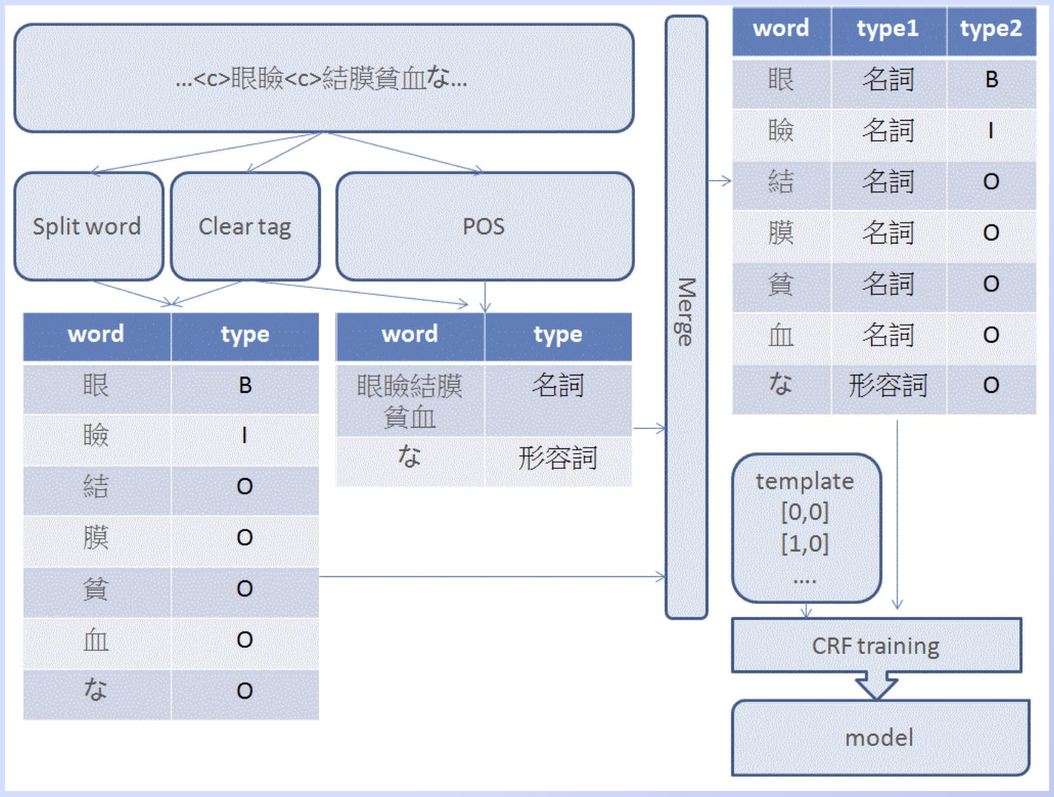


Figure 1: Structure of training processing

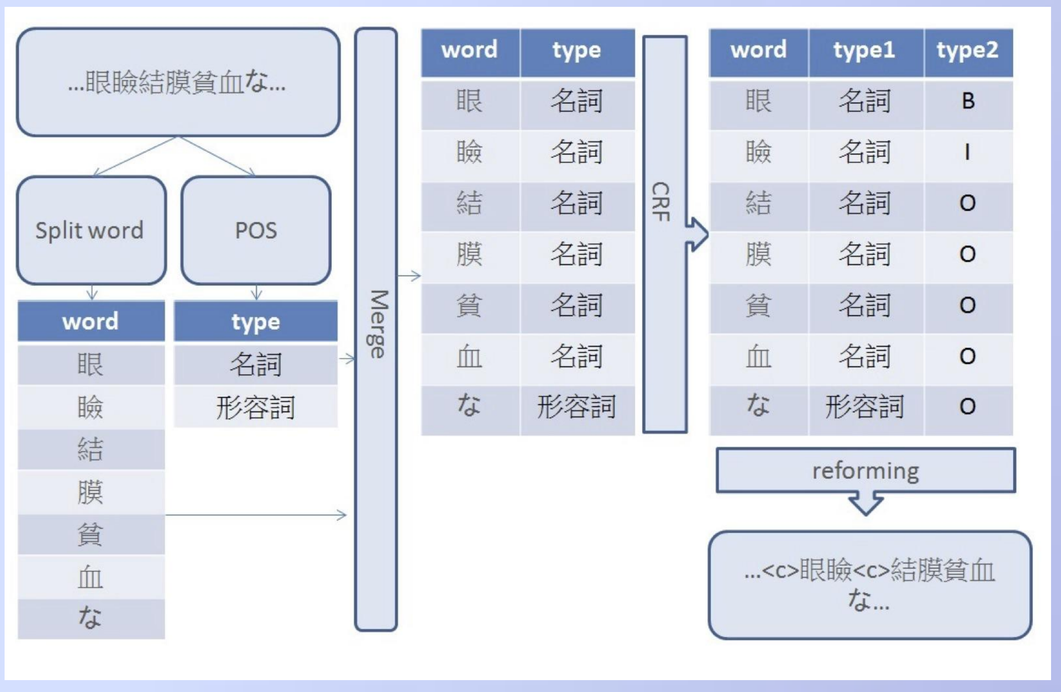


Figure 2: Structure of analysis processing

## Official Results

- Run 01: Combine two dicts and the data based on vocabulary.
- Run 02: Combine one dicts and the data based on words.
- Run 03: Combine two dicts and the data based on words.

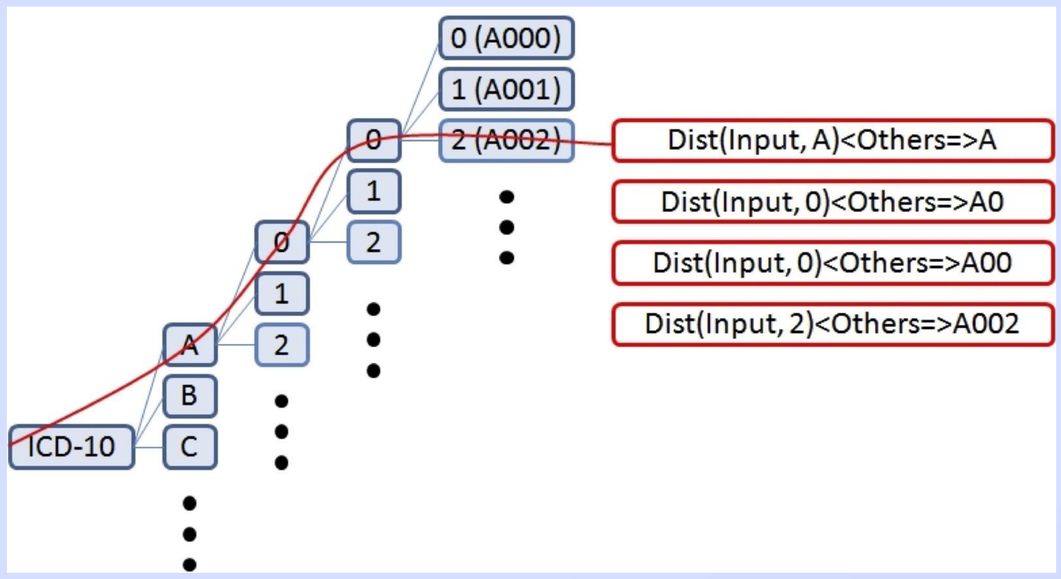


Figure 3: ICD10 tree classifier

Table 1: Official evaluation result with Run-01

| Tag | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 62.68 | 48.62 | 54.76 |
| Family | 37.21 | 76.19 | 50.00 |
| Negation | 51.86 | 51.57 | 51.71 |
| Suspicion | 6.56 | 7.27 | 6.90 |

Table 2: Official evaluation result with Run-02

| Tag | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 62.50 | 47.87 | 54.21 |
| Family | 37.21 | 76.19 | 50.00 |
| Negation | 51.28 | 51.42 | 51.35 |
| Suspicion | 6.67 | 7.27 | 6.96 |

Table 3: Official evaluation result with Run-03

| Tag | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 51.13 | 38.89 | 44.18 |
| Family | 40.32 | 59.52 | 48.08 |
| Negation | 51.37 | 50.71 | 51.04 |
| Suspicion | 8.51 | 7.27 | 7.84 |

## Conclusion

In this time, we try to use three methods to solve the problem, help to find the correct condition name in Japanese. In this three methods, we combine one or two dictionaries into our training model to wish to rise the accuracy rate when the system analysis the raw context. In this three runs, the Run-01 and Run-02 has the close accuracy rate. Use this system to test the past test data. We had the 80% accuracy rate, in this time, the accuracy rate is lower than 70%. According the research, the type of data and amount of data will affect the accuracy rate. If we can have more information data to become our training model. The system can detect more information when it analysis the raw context. And it will have the higher accuracy rate than now. And the complex rate of training model have the positive effect in accuracy rate. The training model have more information that can make the higher accuracy rate.