UE-UD at NTCIR-12 MedNLPDoc Task

Paulo Quaresma
University of Evora
pq@di.uevora.pt

Nga Tran Anh Hang
University of Evora
ex13844@alunos.uevora.pt

ABSTRACT

Technology is the tool that is being used in the various sectors of life and medical is one of them. Electronic medical records (EMR) are now widely used instead of physical documents.

This paper aims to achieve continuing challenges of MedNLP task series in NTCIR-10 and 11. In these tasks, it already attempted named entity recognition (NER) and evaluated the term normalization technology from medical reports written in Japanese, whereas, this task are more advantage, practical and closer to reality application for the medical industry. This task divided into 2 subtasks: (Task1) Phenotyping task requires giving a standard disease names from given medical records, (Task2) creative task to make up ideas to utilize resulting products in the real world. This paper focuses on using tag of speech and improve NER to correctly get sequences of words string in order to achieve the ICD. The experimental result has not shown quite high performance (precision major: 9.6%, recall major: 4.4%, F-measure major: 6.0%). However, it strongly shows a promising result from an international non-speaking Japanese group.

Keywords
Medical records, electronic medical records (EMR), MedNLP, named entity recognition (NER), tag of speech, ICD

Team Name

UE-UD

Subtasks:

(1) Task1 (Phenotyping task)

1. INTRODUCTION

There is no doubt about electronic medical records are replacing paper documents since the importance of technical development for analyzing given information increases rapidly. Therefore, the needs of applying communication technology in medical areas are strongly increasing by years.

To improve and support practical tools for the medical field in the future is one of the main goals of this project.

To process a significant number of documents and gain information about it costs lots of time and workforce. To solve that problem, we apply natural language processing in the medical document to automatically gain information as well as analysis them efficiently.

The main methodology relied on a classical method, which was applied NTCIR-10 and 11.

In the Phenotyping task, by applying tag of speech and named entity recognition to get the sequence of related words as known as keywords and search in ICD-10 and research sequence of keywords in training data to get the best match ICD.

The experiment result has shown a low performance (precision major: 9.6%, recall major: 4.4%, F-measure major: 6.0%) because common diseases matching the sequence of

keywords more than a correct disease, this leads to the low precision and recall. To reduce that mistake, group improve searching engine but strictly select ICD based on repetitive words references ICD description and training data.

This paper analysis the mistake leading to low result and method to improve performance. In addition, it shows evaluation method, future work, and conclusion.

2. TASK & MATERIALS

1.1 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. ICD 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 a 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.

1.2 Data Example

Table 2.1 presents data formats of training data and testing data as following.

File name	Available formats
Training data	172.7kB
Testing data	137.5kB

Table 2.1. Data formats table

Training data: 200 samples

Testing data: 78 samples

(a) Input

<data id="11" sex="FEMALE" age="21"><text type="既往歴">
アレルギー性鼻炎。
１５〜１７歳；円形脱毛症。
１８歳；拒食。
１９歳；うつ、引きこもり、過食。
</text><text type="家族歴">
祖父；高血圧。
父；会社員、母；患者が幼少の頃より多忙。
</text><text type="現病歴">
中学２年生（１５歳）から高校２年生（１７歳）まで原因不明の円形脱毛症の既往あり。
高校３年生（１８歳）、部活のストレスや友人とのケンカが原因で拒食気味になった。
周囲の学力が高いことにより劣等感をもつようになる。
大学に入学後も何をやって良いか目的がわからなくなり、自室に引きこもりがちになる。
自室にて一日中食べている状態が続き、太ったため外出ができなくなった。
食べたものを吐くことはなかった。
一日中自室で横になり食べ続ける状態となって、２００４年６月に当科外来受診。
うつ、心因性摂食障害にてフォローされてきたが、１２月頃より自殺念慮が強くなり、年末に大量服薬により自殺未遂を起こして入院している。
その後も自殺念慮が強いため、加療目的にて入院となる。
</text><text type="現在の愁訴">
自殺念慮</text><text type="入院時現症">
左手首に刃物による切り傷の跡あり。
</text><text type="入院後経過">
入院時、入眠困難のため、経口薬剤にて睡眠コントロールを図る。
加えて、本人の訴え（話し）を傾聴することより薬剤なしでも入眠できるようになる。
入院後は過食行動はみられず、適宜外食を許可し経過観察をしたところ、摂食障害、うつ状態ともに改善してきた。
しかし、退院後や将来のことに対する不安は残っている。
今後は外来フォローということで退院となった。
</text>

(b) Output

<icd code="F508"/>
<icd code="F329"/>
Medical Information
System Development Center (MEDIS-DC) in Japan
and corresponding with ICD-10.<icd code="Z915"/>
<icd code="F510"/>
<icd code="F411"/>

Figure 2.2 Coding example

In the example , given patients current symptoms, historical record, family disease record and testing lab result. Based on those information, the correct ICD match are F508, F329, Z951, F510, F411.

3. METHODS

3.1 Proposed architecture

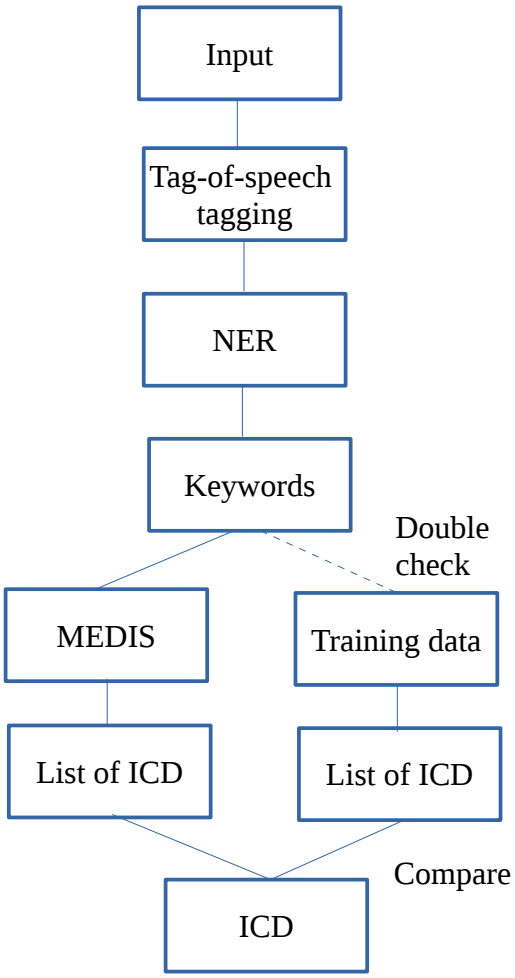

Figure 3.1 Proposed architecture

The outline method of the design illustrates as the figure 2 below. The architecture design is divided into four main processes as
(1) Using a Japanese morphological analyzer – Kuromoji[1], an open source application supporting part-of-speech tagging to assign word categories such as nouns, verbs, particles adjectives,...etc. To extract named entities, we utilize Conditional Random Fileds (CRFs) (Lafferty et al.,2001), undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes[2], in order to gain optimal performance.

(2) To match ICD-10 codes on medical records, we calculated string similarity between given keywords we found in step (1) and all diseases in Hyo Hyun-Byoumei Master, a medical dictionary published by Medical Information System Development Center (MEDIS-DC)[3]. For each disease name in the MEDIS-DC as a corresponding ICD-10 code.

(3) Double checking the matching ICD-10 codes by using training data sets. By applying edit distance similarity in both training and testing data, we scanning all keywords in training data sets compared the best match in target words and get the best fit for each ICD-10 code.

(4) Summing result from (2) and (3) fulfill it in final ICD list

1.3 Evaluation method

The result of Phenotyping task as shown as the following table.

Project	Precision major	Recall major	F major
Ver1	3.2%	7.2%	4.4%
Ver2	9.6%	4.4%	6.0%

Table 3.1 Phenotyping task performance

Performance of the coding task was assessed

using the F-score (β=1), precision, and recall [4].

The performance is not quite good for the first version project because of the common disease terminology occurs in the sequence of keywords that leads to incorrect ICD. To solve the problem, we use training data and MEDIS as references to select correct ICD. As shown in version two, the result is improving.

Compared to other work, this is not the best performance but it is a potential work and we strongly believe that, in the near future this work can be done with a better result.

4. FUTURE WORK AND CONCLUSION

In the Phenotyping task, it is difficult to match ICD in a high accuracy. By using an exact matching method based on ICD-10 and MEDIS with the reference of training data, the performance increases but it is still far away from a final approach to applying in the medical industry. To make this is happening, this project have to mainly focus on matching methodology correctly.

Moreover, once it improves the result, we can think in a further direction, for example, each patient has an ID code and all of the changes, historical disease records, medical treatments all stored as a patient database ID code. In that way, doctors and patients both can access their medication history to check their health. From my opinion, there will be no way to stop technology develop in medical fields.

5. REFERENCES

[1] Kuromoji: Kuromoji is licensed under the Apache License v2.0, http://www.atilika.org/

[2] Andrew McCallum, Wei Li, University of Massachusetts Amherst. E*arly Result for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons.* Amherst, Massachusetts.

[3] MEDIS: ICD Taiou Hyoujun-Byoumei Masters version 3.13, http://www.dis.h.u-tokyo.ac.jp/byomei/index.html (2013).

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