NIL: Simple approach to find the ICD codes

Masao Ito
Nil Software Corp.
nil@nil.co.jp

Masaya Kato
Nil Software Corp.
kato@nil.co.jp

ABSTRACT

To do the task of NTCIR-12 MedNLPDoc, we adopted the simple approach that is match the noun in the medical records with the disease name in the ICD code after morphological analysis of text of medical records. Of course, the effectiveness of this approach is restrictive. But this is our first try and we consolidate the problems and advance toward the next step.

Team Name NIL

Subtask Task-1: Phenotyping task

Keywords

Medical records, ICD-10, MedNLPDoc

1. INTRODUCTION

We use the simple approach to achieve this task. The naive approach means that we extract nouns from the text of medical records and keyword matching with the ICD. There are two reasons for this. One is we want to find the suitable approach for this task, and we think that the plain way gives us the problems clearly. Another is the characteristic of the Japanese disease name; sometimes, it is the Chinese-like sentence. For example, ki-kan-shi-zen-soku-gatsu-pei-nin-shin.

We understand there are many other ways, but this year, it is the first try for us. And we use this simple way.

2. METHOD

Our approach has roughly three steps; first of all, we make the morphological analysis for the text of the medical records by using the JUMAN[1] tool, then (b) try to find out the nouns that might be the candidates as the disease name, finally (c) check the ICD classification that match the candidate word.

In extracting nouns from the result of the morphological analysis, we focus on the list of words that include the medical relating term, such as 'poison ((chuu-doku))[2]', 'incompetence ((fu-zen))', 'disorder ((sho-gai))' and so on. We create this list while experimenting our approach.

When matching the candidate word with the ICD code, we use our rules created for this task. We show some of them:

(1) Addition

(1-1) Just add a qualifier Q:

To match ICD code, the term information might be important. That is, it lasts for a long time or not.

C := QN

To match ICD code, the term information might be important. That is, it lasts for a long time or not.

e.g. chronic ((man-sei)) / acute ((kyuu-sei))

(2) Synthesis

(2-1) region

If we find out the region of the patient body in the text and a keyword of disease, we generate the synthesized word.

C := QN or Q'N

Q means segment indicates simple body region that has not detailed keyword

e.g. N: "malignant neoplasm ((ga-n))", Q: "metastasis to right lung ((migi-hai-ten-i))"

Q might have the right and left information, and we extract this and synthesize.

Q': "metastasis to lung ((hai-ten-i))".

We can get the new candidate word "metastaic lung cancer ((ten-i-sei-haigan))"

(2-2) drug

C := QN

Q: segment indicates the relation to drug (e.g. "drug-induced ((yaku-zai-sei, yaku-butsu-sei))")

N: the word is "disorder ((sho-gai))" or "inflammation((en))"

First, we remove the word indicating the degree of symptom, such as severe ((jyuu-do)) and the name "function ((ki-nou))".

If we find the words "drug-induced" and "disorder of liver function", we create the candidate word, "drug induced hepatitis ((yaku-butsu-sei-kan-sho-gai))". We can find the similar case such as "toxic ((chuu-doku-sei))".

Figure 1 coding rule (excerpt)

(C means candidate word to get the ICD code. N is the extracted noun, Q means Qualifier)

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman. html

[2] The double brackets "(())" mean the Japanese reading of a Kanji. A hyphen indicates an end of a kanji.

(2-3) age and fracture

We use information of the patient's age.

C:= QN_2N_1 or QN_1N_2

Q: the segment indicates the age of the patient.

N_1: the word is "fracture ((kotsu-setsu))" or "osteporosis ((kotsu-so-sou-shou))".

N_2: region of fracture

If the patient is older than 65 year's old and brake a bone, the candidate term is "senile osteoporosis ((rou-nen-sei-kotsu-so-sou-shou))". If N_2 indicates the region of the patient, add it.

e.g. if N_1 is "osteporosis" and the patient is younger than 65 from the age property, we choose the "juvenile osteoporosis ((jyaku-nen-sei-kotsu-so-sho-shou))".

(3) Change

Replace the qualifier to other similar word: Q_xN -> C := Q_yN

We use this rule in order to match the ICD-code description.

e.g. senile ((ro-jin-sei)) -> ((ro-nen-ki))

Figure 2 coding rule (excerpt) cont'd

(C means candidate word to get the ICD code. N is the extracted noun, Q means Qualifier)

3. RESULTS

In our approach, the results applying to training data are good, but in case of the test data set, it is hard to get the correct answer. Because the training data set has the diagnosis of disease, but the test data has not. We have to add the function to aid the diagnosis.

As for id 0 of training data, we get the ICD code, J30.4 and O99.5 in our program. The original numbers coded are O34.3, O99.5, J30.4 and Z35.1. To obtain O34.3 is easy, but our program doesn't correctly extract the number (simple programming error). On the other hand, to get the Z35.1 is difficult. Because we estimate the past (two) abortion(s) from three pregnancies and one delivery. This time we just check the words in a literal manner.

Then we check id 0 of test data set. Our program produces the several words; B169,B029,R53,R509,R060 and R234. But we cannot lead the primary disease (e.g. B24) because absence of description.

4. CONCLUSION

Our approach is simple approach just focusing on the keyword search with the manipulation on the word such as adding other words and replacing words. We believe this approach is suitable for encoding to ICD-10 when the name of the disease is in the text. But this approach cannot estimate the disease, a matter of course.

But currently we try to do the other approach. In this approach, we translate the several keywords into English, and match with the SNOMED-CT. The SNOMED-CT has the mapper to ICD-10, so we can go to get the ICD-10 code through SNOMED-CT.

Because SNOMED-CT has a kind of semantic network. And if we can properly trace the network, we think we can get the right ICD-10 code without the name of disease in the text. If we have the opportunity, next time we hope to report the results.