The OKPU System in NTCIR11 MedNLP2: An IR Approach to ICD-10 Code Identification

Genichiro KIKUI Yasuhiro TAJIMA
Okayama Prefectural University
Kuboki 1-1-1, Soja-shi, Okayama, 719-1197, JAPAN
{kikui, tajima}@cse.oka-pu.ac.jp

ABSTRACT

This paper describes an IR (Information Retrieval) approach to identifying the ICD-10 code of a medical term, such as a disease name or a description of a symptom or a complaint), in a medical text. In this approach, we prepare a *dictionary* of disease names, each paired with a corresponding ICD-10 code(s). The system searches for the disease name most relevant to the input, and returns the ICD-10 code paired with the disease name in the dictionary. In IR terms, disease name in the dictionary can be regarded as a *document* and an input medical term as a *query*. In order to handle an input which does not exactly match with any disease names in the database, we introduce two kinds of partial matching and a context search, where a query includes context words of the input term. Preliminary evaluation for the MedNLP2 test set shows that with this simple approach our system correctly identified 54% of the input medical terms.

Team Name

OKPU

Subtask

Task2 (ICD-10 code identification)

Keywords

Medical terms, Information Retrieval, sense identification

1. INTRODUCTION

The task (2) of MedNLP-2[1], hereafter, the task-2, is essentially sense identification of a linguistic expression. Sense identification task can be generally classified into two types: one is to map a given linguistic expression into some extra-linguistic symbol(s)/label(s) that represents the meaning (hereafter, semantic symbols). The other is to relate a given linguistic expression to different but semantically equivalent (linguistic) expressions. The task-2 falls into the first type.

This task requires a database, or a dictionary, that defines the meaning of each semantic symbol, which may be given by natural language texts or simply a list of linguistic expressions corresponding to the symbol. The system searches for the symbol whose definition best matches with the linguistic expression in question including its surrounding context. There are at least two problems in semantic identification. The first problem is how to handle a variety of surface forms that corresponds to the same semantic symbol. The second problem is how to handle ambiguities, where the same surface form can corresponds to different semantic symbols.

In this work, we take an IR (Information Retrieval) approach, where the system searches the dictionary database for the medical term or description most relevant to the input. Then, the system just returns the ICD-10 code linked with the retrieved term.

In what follows, we describe our system in Section 2, then shows preliminary evaluation and discussion, followed by a summary.

2. SYSTEM DESCRIPTION

2.1 Overview

An overview diagram of our system is shown in Figure 1. At first, the query formulation module converts an input medical term, together with its surrounding context, into a search query. Then, the basic retrieval module searches the dictionary database for terms which shares at least one character with the query. Finally, the filtering module chooses the most appropriate candidate applying exact match (EXACT), character-based partial match (CBPM), and sub-term-based partial match (SBPM).

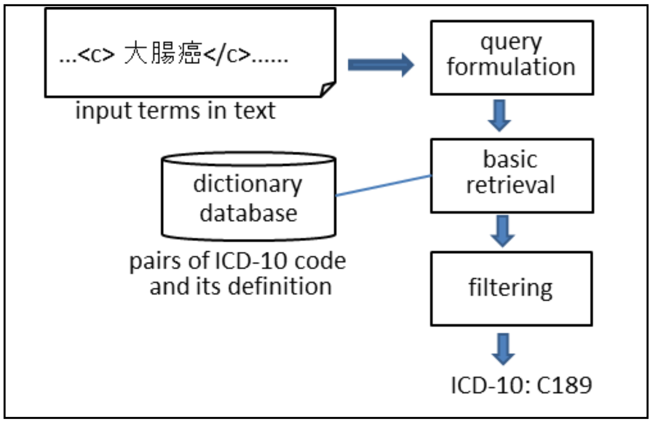

Figure 1: Overview Diagram of the System

2.2 Dictionary Database of ICD 10 Codes

Before describing modules, we explain the dictionary database. The dictionary database is a collection of records. Each record comprises of a medical term, called *the text part*, and the corresponding ICD-10 code(s), called *the ICD part*. A medical term includes a disease name, symptom, complaint, etc. An ICD-10 code consists of an alphabet character followed by two or three numeric characters, such as ‘E11’ or ‘E323’. Note that we limit the numeric part of an ICD-10 code to at most three digits by deleting the fourth digits if exists.

The dictionary database used in this work is created from MEDIS Standard Masters (ICD Hyoujun-Byoumei Master)[2] and from Life Science Dictionary[3]. Concretely speaking, a database record is filled with an ‘index term’ (i.e., medical expression) of the Standard Masters for the text part and its corresponding ICD-10 code for the part of ICD-10 code. The resulting database contains 95,206 records.

2.3 Query Formulation

This module formulates a query for the basic retrieval. Since Task-2 presupposes the medical term in question is correctly located in the input text (parenthesized with <c> and </c>), the simplest strategy is to just use the term without any context. For example, for the input text as follows:

胸痛時頚部、顎下部に<c>放散痛</c>あり。

we the query string is “放散痛”.

When the basic retrieval could not find any records in the dictionary database, the module uses the entire line (i.e., the region separated by linefeed characters) containing the term. For the above example, the query string is “胸痛時頚部、顎下部に放散痛あり”. This type of query is called “Context-based Query”.

2.4 Basic Retrieval

This step is responsible for obtaining initial set of database records. In order to keep recalls, the system searches for database records whose text part shares at least two adjacent characters (i.e., a bigram of characters) with the input. This is implemented with Apache-solr[4] by setting options to using character bigram tokenizer and default ranking formula based on the tf-idf scoring function.

2.5 Filtering

2.5.1 Exact Match

Exact match (EXACT) is the simplest but very powerful approach when a comprehensive dictionary database is available. The exact match filter simply chooses records whose text part is identical to the input medical term and returns its ICD-10 part.

When more than one ICD codes are obtained (i.e., the database has multiple records with the identical description but different ICD-codes), the system randomly chooses one candidate. For the development set, we found 7% of exactly matched records fell into this case.

Our preliminary evaluation has shown that the exact match achieved precision and recall of 0.83 and 0.38, respectively, for the development set (shown in Table 2).

2.5.2 Partial Match Filters

The exact match filter as described above can handle only 45% inputs of the training set. In order to handle the remaining, we applied two partial match filters: PARTIAL-1 and PARTIAL-2. The system first applies PARTIAL-1 to the original candidates (obtained by the basic retrieval). If the system can chose one at least one candidate, then the filtering step terminates. Otherwise, the system tries PARTIAL-2.

2.5.2.1 PARTIAL- 1 (prefix/suffix match)

This filter tries to select a record whose text part is an affix (prefix or a suffix) of the input. This filter also allows the reverse case, where the input is an affix of the text part of a database record. In order to choose the best candidate, this filter calculates the score of the matched record as follows:

Score = 1/|len(desc)-len(input)|

As shown, the score is the inverse of the number of unmatched characters. Note that this filter is applied to an input where its candidates does not include exact match, thus the denominator is guaranteed to be a positive number.

2.5.2.2 PARTIAL-2 (feature-based match)

When a medical term consists of multiple sub-term units, each sub-term has a role, called a *feature*, in the entire term. For example,

“開放性胸部気管損傷(open chest tracheal injury)”,

which can be divided into three parts : “開放性(open)”, “胸部(chest)” and “気管損傷(tracheal injury)”. They respectively correspond to “manner”, “body position” and “core” features. These features, except the core-feature, are automatically extracted by using pattern matching since terms or sub-terms for a each feature has specific suffixes or patterns as shown in Table 1. The ‘core feature’ is the remaining part, which normally occupies the rightmost position of the term. Each feature is assigned a heuristically determined weight as shown in Table 1. Note that the core-role-part has a very large weight.

The PARTIAL 2 filter first extracts features from both the input term and the text part of the database record. Then, it calculates the matching score by summing up weights of matched features. It finally chooses the record with the highest matching score. When the largest score among the candidates is below the predetermined threshold, this filter makes no output..

Table 1: Roles of Subterm

Feature name	type	manner	body position	core
Pattern	.*型	.*性	.*部	-
Example	I 型 Type-I	開放性 open	胸部 chest	気管損傷 Tracheal injury
Weight	1	1	2	10

3. Evaluation

We evaluated our method with the development set and the test set provided by the organizers. For these two data sets, we tried two methods as shown below:

Method-1: The above mentioned entire steps with the threshold of PARTIAL-2 to 9.0, which means that the core-role should be idential.

Method-2: The entire steps with the threshold of the PARTIAL-2 to 0.0, which means that the partial match-2 is used just for re-ranking the candidates.

3.1 Results for the Development Set

Evaluation results for the development set using Method-1 and Method-2 are shown in Table 2 and Table 3 respectively. EXACT, PARTIAL-1 and PARTIAL-2 in the tables correspond to exact match, partial match with affix, and feature-based partial match, respectively. OK means that that an output of the system is identical (i.e., has the same character string) to the gold standard provided by the organizers.

As far as these tables are concerned, exact match is effective. It covers 45% of input terms with more than 82% precision. As compared with exact match, partial match is not so effective. In fact, results of partial match 2, where we tried to use the structural information of a term, were disappointing.

Table 2: Method-1, Development set

	OK (rat7e)	NG	TOTAL
EXACT	1245 (0.828)	258	1503
PARTIAL-1	122 (0.722)	47	169
PARTIAL-2	56 (0.218)	201	257
CONTEXT	266 (0.199)	1068	1334
NORESULTS	0 (0.0)	41	41
TOTAL	1689 (0.511)	1615	3304

Table 3: Method-2 Development set

	OK (Rate)	NG	TOTAL
EXACT	1245 (0.828)	258	1503
PARTIAL-1	122 (0.722)	47	169
PARTIAL-2	439 (0.328)	898	1337
CONTEXT	5 (0.018)	269	274
NO RESULTS	0 (0.0)	21	21
TOTAL	1811 (0.548)	1493	3304

.

3.2 Results for the Test Set

3.2.1 Submitted Results

Evaluation results for the test set using Mthod-1 and Method-2 are shown in Table 4 and 5. As shown, overall accuracy of Method-2 was 39%, about 15 points lower than that of the development set. The main reason comes from an error (bug) in forming ICD-10 codes (which omitted '_' characters)..

Table 4: Method-1, test set

	OK	NG	TOTAL
EXACT	571 (0.627)	339	910
PARTIAL-1	107 (0.413)	152	259
PARTIAL-2	23 (0.277)	60	83
CONTEXT	94 (0.109)	768	862
NO RESULTS	0 (0.0)	22	22
TOTAL	795 (0.372)	1341	2136

Table 5: Method-2, Test set

	OK	NG	TOTAL
EXACT	571 (0.627)	339	910
PARTIAL-1	107 (0.413)	152	259
PARTIAL-2	164 (0.211)	612	776
CONTEXT	5 (0.028)	175	180
No Results	0 (0.0)	11	11
TOTAL	847 (0.397)	1289	2136

3.2.2 Revised Results

Results shown in Table 4 and Table 5 include trivial errors which are irrelevant to algorithms and methods. Thus we introduce the revised results in Table 6 and Table 7. These results were generated by just adding '_' to each output (in the submitted version) whose numeric part was less than 3 digits.

Table 6: Method-1, test set

	OK	NG	TOTAL
EXACT	792 (0.870)	118	910
PARTIAL-1	155 (0.598)	104	259
PARTIAL-2	32 (0.385)	51	83
CONTEXT	109 (0.126)	753	862
NO RESULTS	0 (0.0)	22	22
TOTAL	1088 (0.509)	1048	2136

Table 7: Method-2, Test set

	OK	NG	TOTAL
EXACT	792 (0.870)	118	910
PARTIAL-1	155 (0.598)	104	259
PARTIAL-2	197 (0.254)	579	776
CONTEXT	5 (0.028)	175	180
No Results	0 (0.0)	11	11
TOTAL	1149 (0.537)	987	2136

As compared with the results for the development set, the test set was much harder. Above all, exact match is effective with 42% recall and 87% precision. Considering Partial match of affixes (PARTIAL-1) is a bit less effective.

A common problem for both development and test sets lies in the wider variations of descriptions that should be mapped into the same ICD code. These varieties include standard and non-standard abbreviations, foreign language terms in original or transliterated spelling, Japanese specific Katakana-Hiragana-Kanji variations, etc.

Another problem is incompleteness of dictionary. This problem is serious because we should always consider different ICD codes for every term in the database.

4. Summary

In this paper, we introduced a simple IR-based approach to ICD-10 code identification of medical diagnosis, symptoms and complaints. Our system correctly identified 54% of ICD-10 codes in the MedNLP 2 task 2 test set.

Future directions include trying various IR techniques including term expansion (considering terminological properties), latent indexing etc. Using hierarchical structure of ICD-10 codes may improve the performance.

Acknowledgement

We are grateful to MEDIS-DC (Medial Information System Development Center) for allowing us to use ICD Masters [3]. We would also appreciate Professor Shuji Kaneko for allowing us to use the life Science Dictionary.

References

[1] Eiji Aramaki and Mizuki Morita: “Overview of the NTCIR-11 MedNLP-2 Task”, NTCIR-11, this volume (2014).

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

[3] Shuji Kaneko et al.: “Life Science Dictionary: a versatile electronic database of medical and biological terms”, Asialex, pp.434-439 (2003) http://lsd.pharm.kyoto-u.ac.jp/ja/index.html

[4] Apache-SOLR: http://lucene.apache.org/solr/ .