Statistical Selector of the Best Multiple ICD-coding Method

Eiji Aramaki\textsuperscript{a}, Takeshi Imai\textsuperscript{a}, Masayuki Kajino\textsuperscript{b}, Kengo Miyo\textsuperscript{a}, Kazuhiko Ohe\textsuperscript{a}

\textsuperscript{a} The University of Tokyo Hospital, Japan
\textsuperscript{b} Japan Research Group for Medical Ontology, Japan

Abstract

The International Classification of Diseases 10th version (ICD-10) is one of the most standard and important disease classifications. Since computerized ICD-10 coding systems have drawn a great deal of attention in the medical field, a great number of different coding systems have been proposed. The present paper proposes a hybrid architecture of different coding systems. First, given an input disease name, three coding systems output codes with their confidence scores. A C4.5-based system selector then selects the best output by using both input statistics and the confidence score from each system. The experimental results demonstrated that the selector significantly boosts the overall performance (+3.4 points).

Keywords:
ICD-10, International Classification of Disease Codes, Coding, Decision Trees, Natural Language Processing.

Introduction

International Classification of Diseases 10th version (ICD-10) \cite{ICD10} was endorsed by the 43rd World Health Assembly in 1990 and came to be used by World Health Organization (WHO) member countries from 1994. ICD-10 has become the international standard diagnostic classification for all general epidemiological purposes and many health management purposes, such as the analysis of the general health circumstances of population groups and monitoring of the incidence and prevalence of diseases \cite{WHOICD10}.

Therefore, many medical doctors or coders (who are the special staff for the ICD coding) have selected a suitable disease code for each patient. As pointed out in a previous study \cite{ICDcoding}, the coding task is an important semantic issue because it requires a very high level of understanding of the meaning of the patient data.

In such a situation, in order to reduce the heavy burden of coding, a number of automatic coding systems/methods have been proposed. For example, systems that output a code from a discharge summary \cite{DischargeSummary}, a coding system for autopsy reports \cite{AutopsyReport}, a coding system for input disease names \cite{DiseaseName}, coding frameworks \cite{Codingsystems} \cite{CodingFrameworks}, and a machine learning coding for a diagnostic text \cite{MachineLearningCoding}, have been proposed. Although there are numerous system variations, the approaches are basically classified into two types.

One approach is to directly capture the relation between an input and an ICD category \cite{DirectCoding} \cite{DirectCodingFrameworks}. In the present study, we refer to such systems as direct systems.

However, direct systems often suffer from complex language phenomena. For example, given the term “leg eczema”, which should be classified into an ICD-10 code [L309] “inflammation of the skin”, in order to realize successful coding, a direct system requires knowledge such as:

\begin{enumerate}
\item “Leg” has “skin”,
\item “Eczema” is-a-kind-of “inflammation”.
\end{enumerate}

In order to obtain such knowledge, most systems utilize an ontology/knowledge-base. However, it is difficult to fully cover such knowledge.

Another approach \cite{ExampleBasedCoding} \cite{MachineLearningCoding} is to retrieve the most similar coding example from a database, and output its code as is. We herein refer to such systems as example-based systems. The example-based system easily realizes the above coding, if the system can find a coding example, as follows:

\begin{enumerate}
\item (Input) “Leg eczema”,
\item (Coding-example) “Foot eczema” \[\rightarrow\] [L309].
\end{enumerate}

In summary, both approaches have different strengths and weaknesses. Therefore, the present paper proposes a hybrid architecture, which selects a suitable system for each input. First, given an input, three coding systems output codes with their confidence scores. Then, a selector based on a C4.5 decision-tree \cite{C45} decides the best output by using information from both the confidence score of each system and input statistics.

The experimental results demonstrated that the selector significantly boosts the overall performance.

It should be noted that previous studies use various input types/forms, such as a discharge summary \cite{DischargeSummary}, an autopsy report \cite{AutopsyReport}, and a disease name \cite{DiseaseName}. In this paper we use a dis-
ease name as a system input, since it is the most informative clue in the coding task.

Although the proposed system does not depend on language, we conducted experiments using disease names in Japanese.

Material

This section describes ICD10 [1]. The ICD10 has a tree structure, in which a leaf code consists of one alphabetical character <A-Z>, three digit numbers <1-9>, and its description. We call the description an ICD-term.

Figure 1 shows an example of an ICD10 code with its ICD term “Cholera, unspecified”. The first alphabetical character <A> indicates a top level category, and the second <0> indicates the second level category. In the same manner, the third and fourth numbers indicate the third and fourth level categories.

Figure 1 Example of ICD-10 code and ICD-term

Methods

The task to be solved is to output a suitable ICD-code (code) for an input disease name (input). The idea of the proposed method is to dynamically select the best output from three systems. We first explain the three coding systems of the present study and then present the system selector.

(1) Direct System

As mentioned above, most direct systems utilize knowledge-base/ontology. However, such resources are not always available in Japanese and other non-English languages. Therefore, the direct system of the present study utilizes ICD-terms, because ICD-terms are available as a knowledge-base in native languages in most countries.

The present direct system captures the similarity between an input and the ICD-terms. First, this system searches all of the ICD-terms, and extracts the ICD-term that is most similar to an input. For calculating the similarity, we used two measures: (1) a BM25 [13] and (2) an edit-distance [14].

BM25: The BM25 \( \text{simBM25} \) is a word-based similarity, which is employed in a state-of-the-art information retrieval system, the Okapi-system [15], which is defined as follows:

\[
\text{simBM25} = \sum_{t \in T} \left( W_d \times W_q \right),
\]

where

\[
W_d = \frac{(k_1 + 1) f_t}{k_1 ((1 - b) + b \times dl/avdl)},
W_q = \log \frac{N - n + 0.5}{n + 0.5}.
\]

In this formula, \( T \) is the set of words appearing in both the input and the ICD-term, \( f_t \) is the number of occurrences of a word \( t \), \( dl \) is the length of an ICD-term, \( avdl \) is the average length of the ICD term, \( N \) is the total number of ICD-terms, \( n \) is the number of extracted ICD-terms, and \( k_1 \) and \( b \) are the constants determined from the preliminary experiments\(^2\). For details, see [13].

Edit-distance: The edit-distance similarity \( \text{simED} \) is a character-based similarity, which is based on the minimum number of point mutations required to change an input into an ICD-term, where a point mutation involves: (1) changing a character, (2) inserting a character, or (3) deleting a character. For details, see [14].

The final similarity is a weighted sum of \( \text{simBM25} \) and \( \text{simED} \), as follows\(^3\). We refer to this similarity as a direct-sim.

\[
direct-sim(input, ICD-term) = W_1 \text{simBM25} + W_2 \text{simED}.
\]

The final output code is decided by the code of the most similar ICD-term.

(2) Example-based System

The example-based system retrieves the most similar coding example from a database (in experiments, we used 15,551 coding examples) and outputs the code of the retrieved example. In addition, in this similarity calculation, a weighted sum of BM25 and the edit-distance was employed. We refer to the similarity between an input and a coding-example as an example-sim.

The final output code is decided “as is” by the code of the most similar example.

(3) Combination System

In preliminary experiments, the example-based system demonstrated high performance in rough classification, judging only the matching of the first and second digits. Considering this example-based advantage, we designed another system, referred to as the combination system.

The combination system first decides the first and second ICD-code using the example-based system. For example, if the example-based-system outputs \([L309]\), then \([L3**]\) is the scope of the following searches. A detailed code (third and fourth digits) in the scope is then decided by a direct system.

These system frameworks are illustrated in Figure 2.

---

\(^1\) Some diseases have one more digit (one alphabetical character and four numbers). Since such codes are rare (7.6%), the present paper does not handle them.

\(^2\) We set \( k_1 = 1.5 \) and \( b = 0.75 \) in the experiments.

\(^3\) We set \( W_1 = 0.9 \) and \( W_2 = 0.1 \) in the experiments.
Statistical Selector

The selector decides the suitable system for each input. Here, the selector uses two types of information (features): (f1) information from each system and (f2) information from an input. The detailed features are described in Table 1.

By using a training-set, the selector learns the relations between the features and which system is correct. In this learning, we utilized C4.5 [12], because it clearly shows the effectiveness of each feature.

Note that if the selector estimates that two or more systems are correct, we give the following priority: EXAMPLE-BASED > COMBINATION > DIRECT. For example, if the selector estimates that both COMBINATION and DIRECT are correct, then a COMBINATION output is employed.

In the same way, if the selector estimates that all systems are incorrect, an EXAMPLE-BASED output is employed. This priority is based on preliminary experimental results.

Table 1 Features for system selector learning

<table>
<thead>
<tr>
<th>(f1)</th>
<th>Max Direct-sim: The maximum direct-sim in the k-best ICD-terms.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min Direct-sim: The minimum direct-sim in the k-best ICD terms.</td>
</tr>
<tr>
<td></td>
<td>Example Code Variations: the number of second level category variations in extracted k coding-examples. For example, given [L309] [L300] and [B100], this value is 2 ([L3**] and [B1**]).</td>
</tr>
<tr>
<td></td>
<td>Max Example-sim: The maximum direct-sim in the k-best coding-examples.</td>
</tr>
<tr>
<td></td>
<td>Min Example-sim: The minimum direct-sim in the k-best coding-examples.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(f2)</th>
<th>Input Length: The number of the input words.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input Headword: The headword (last word) of an input. (We assume a head-final principle).</td>
</tr>
<tr>
<td></td>
<td>Dictionary: Whether the input is found in the dictionary entry.</td>
</tr>
</tbody>
</table>

Experiments

Experimental Settings

To investigate the performance, we used two materials: (1) a set of coding examples [16] consisting of 15,551 pairs (Japanese disease name and its corresponding ICD code), and (2) 995 disease names, which are input in several Japan hospitals. We divided (2) into two sets as follows: a training-set (for selector training) and a test-set (for evaluation).

Both sets were annotated with their correct codes (gold standard). If the gold standard code is ambiguous for only the disease name, we removed it. The number of codes in each set is shown in Table 2.

To acquire the Japanese word units in a term, we used a morphological analyzer developed in a previous study [17].

Table 2 Number of coding examples

<table>
<thead>
<tr>
<th></th>
<th># of codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding-example</td>
<td>15,551</td>
</tr>
<tr>
<td>Training-set</td>
<td>495</td>
</tr>
<tr>
<td>Test-set</td>
<td>500</td>
</tr>
</tbody>
</table>

In the evaluations, we used three measures (1-3) and compared the following four systems (a-d), as follows:

(1) EXACT: The exact (three digits) match.
(2) ROUGH: Only the first and second digits match, e.g., [L309] and [L309].
(3) 10-BEST: The system outputs 10 candidates, which are judged as to whether they contain the gold standard.
(a) DIRECT: The direct system.
(b) EXAMPLE: The example-based system.
(c) COMBINATION: The combination system.
(d) PROPOSED: The proposed system, selecting the best from the above three systems.

Results

The results are shown in Table 3. For two measures, PROPOSED showed the highest accuracy. In particular, in EXACT, the proposed system demonstrated significantly higher accuracy than the other systems in the McNemar test [18] (p = 0.05).

On the other hand, the ROUGH accuracy of the PROPOSED system is slightly (although not significantly) lower than EXAMPLE-BASED system. One reason for this is that the selector performed training based on the EXACT metrics.

Table 3 Results

<table>
<thead>
<tr>
<th></th>
<th>(1) EXACT</th>
<th>(2) ROUGH</th>
<th>(3) 10BEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) DIRECT</td>
<td>33.6% (168)</td>
<td>52.8% (264)</td>
<td>61.2% (306)</td>
</tr>
<tr>
<td>(b) EXAMPLE-BASED</td>
<td>63.6% (318)</td>
<td>81.6% (408)</td>
<td>65.0% (352)</td>
</tr>
<tr>
<td>(c) COMBINATION</td>
<td>52.4% (262)</td>
<td>74.2% (371)</td>
<td>81.6% (408)</td>
</tr>
<tr>
<td>(d) PROPOSED</td>
<td><strong>67.0% (335)</strong></td>
<td>80.4% (402)</td>
<td><strong>83.2% (416)</strong></td>
</tr>
</tbody>
</table>

* Numbers in bracket indicate the number of the correct outputs. The underlined value in (d) indicates significantly higher values than (a), (b) and (c).

4 In experiments, we set k=10.

5 A word boundary in Japanese language is not explicit.
Table 4 Frequencies of appearance of each system in the proposed hybrid system (PROPOSED)

<table>
<thead>
<tr>
<th>Selected frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRECT</td>
</tr>
<tr>
<td>EXAMPLE-BASED</td>
</tr>
<tr>
<td>COMBINATION</td>
</tr>
</tbody>
</table>

Table 5 Output examples

<table>
<thead>
<tr>
<th>Input disease name</th>
<th>Gold standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCV antibody test positive [DIRECT]</td>
<td>(R768) Immunological injury of serum (a correct code &amp; its ICD term)</td>
</tr>
<tr>
<td>HBc antibody test positive [EXAMPLE-BASED]</td>
<td>[R768] Immunological injury of serum</td>
</tr>
<tr>
<td>E incompatibility [EXAMPLE-BASED]</td>
<td>[O360] Maternal care for RH allogenic immunity</td>
</tr>
<tr>
<td>D incompatibility [O360]</td>
<td>[A491] Chain coccus infection</td>
</tr>
<tr>
<td>Hemolytic streptococcus infection [COMBINATION]</td>
<td>[A491] Chain coccus infection</td>
</tr>
<tr>
<td>Coliform bacillus infection [A491]</td>
<td>[A305] Leper mass type lepra</td>
</tr>
<tr>
<td>Type-LL Hansen's disease [COMBINATION]</td>
<td>[A305] Leper mass type lepra</td>
</tr>
<tr>
<td>Type-1 Hansen's disease [A300]</td>
<td>[A305] Leper mass type lepra</td>
</tr>
<tr>
<td>Palpebra tumor mass [DIRECT]</td>
<td>[K130] Lips of mouth disease</td>
</tr>
<tr>
<td>Lips of mouth sore [DIRECT]</td>
<td>[K130] Lips of mouth disease</td>
</tr>
</tbody>
</table>

* In the case that DIRECT is selected, a coding example is not used.

Table 4 shows the frequency of each selected system, and Table 5 shows some output examples.

Discussion

We then investigated an important clue for the selector. In a decision tree produced by the selector, the first branch is “if the max example-sim > X, use EXAMPLE-BASED”, and the second branch is “if example code variation < X, use EXAMPLE-BASED”. Based on this observation, we can see that the selector provided several opportunities for the EXAMPLE-BASED system.

For more detailed discussions, we investigated each system performance using the following measures:

\[ \text{Precision} = \frac{\# \text{ of correct outputs}}{\# \text{ of outputs}} \]

\[ \text{Recall} = \frac{\# \text{ of correct outputs}}{\# \text{ of test-set}} \]

In this investigation, a system will reject an output if the similarity is less than various thresholds. We examined two systems: (1) the DIRECT system (which relies only on direct-sim) and (2) the EXAMPLE-BASED system (which relies only on example-sim) (Figure 3).

As shown in the figure, EXAMPLE-BASED generally showed higher performance, but in some parts (top-left in the figure) DIRECT is better than EXAMPLE-BASED. Considering the potential of EXAMPLE-BASED, the strategy of the selector, which gives the first chance to EXAMPLE-BASED, is reasonable.

Finally, this figure also shows that we can freely control the system precision and recall. This can be a strong advantage from a practical point of view.

Related Studies

Thus far, a number of studies have investigated automatic coding [4-11, 19, 20, 22-25]. Most of these studies have focused on methods for knowledge representation such as a semantic frame-based ICD representation [9], a conceptual graph-based representation [5], a semantic markup [10], and an analytical index [19]. It is difficult to compare these systems to that of the present study for three reasons: (1) their inputs were different from ours, (2) we dealt with the Japanese language, and (3) more importantly, the present study focuses on the hybridization of systems. We can expect that the present system will have higher performance by incorporating the advantages of their methods.

Another important aspect of coding studies is the coding algorithm such as frequency statistics [20], a Naïve-bayes classifier [11], and an open registry algorithms [22]. Although various approaches have been proposed, each system has strengths and limitations, as pointed out in a study [23], motivating the proposed hybrid approach.

Note that some studies focused on a limited domain. For example, Tagliabue[22] dealt with a cancer domain, and Schnitzer [23] concentrated on a child injury domain. Such a limited domain enables a more finely-tuned system to be developed, leading to higher performance. In the future, we intend to incorporate such domain-limited systems into the proposed hybrid framework.

At present, automatic coding accuracy systems have not yet been perfected. However, Hohnloser [4] reported that their coding system improves human coding quality. In addition, several systems have been investigated in hospitals for years ([24]: five years, [25]: one year). In the future, we intend to examine the performance of the proposed system in a real-world environment.

Conclusion

The present paper proposed a hybrid architecture of different coding systems. First, given an input disease name, three coding systems output codes with their scores. A C4.5-based se-
Acknowledgments

This research was supported in part by a Grant-in-Aid for Scientific Research through the Japan Society for the Promotion of Science (Project Number: 16200939, F.Y.2004-2007 and 18700133, F.Y.2006-2007) and by the Research Collaboration Project with Japan Anatomy Laboratory Co. Ltd.

References


Address for correspondence

Eiji Aramaki, The University of Tokyo Hospital;7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan +81-3-5802-7115;aramaki@hcc.h.u-tokyo.ac.jp