MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190536

An Information Retrieval Approach to ICD-10 Classification

Hee Park, José Castaño, Pilar Ávila, David Pérez, Hernán Berinsky, Laura Gambarte, Daniel Luna, Carlos Otero

Departamento de Informática en Salud, Hospital Italiano de Buenos Aires, Argentina

Abstract

ICD-10 (International Classification of Diseases 10th revision) is a classification code for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. This paper describes an automatic information retrieval approach to map free-text disease descriptions to ICD-10 codes. We use the Hospital Italiano de Buenos Aires (HIBA) terminology data mapped to ICD-10 codes as indexed data to find an appropriate ICD-10 code using search engine similarity metrics.

Keywords:

Clinical Coding, International Classification of Diseases, Natural Language Processing

Introduction

The International Classification of Diseases code [1] is a very useful tool in a wide range of medical processes including statistical analysis of morbidity rate and mortality rate, medical reimbursement and resource allocation [2]. The registry by medical practitioners needs additional training time and manual encoding is not a simple task because not every health professional records in the same way [3].

Automated coding of ICD-10 is not an easy string matching work due to patient context's dependency [4]. There are two approaches to automated coding of medical texts to ICD-10: dictionary projection and supervised machine learning methods or the combination of both [3,4,5,7,8,9,10]. Dictionary projection methods use string matching techniques: n-grams, edit-distance, synonyms and abbreviations expansion, stop-word elimination, frequency of use, spell checkers and stemming. Machine learning methods include Support Vector Machine (SVM), Naive Bayes, Maximum Entropy Modeling (MaxEnt) and Gradient Boosting (XGBoost) [6].

The Hospital Italiano de Buenos Aires (HIBA) has a Spanish interface vocabulary where each term is mapped to SNOMED CT as its reference vocabulary. It has 207,000 post-coordinated concepts in its terminology system. This system provides services to the Hospital Italiano healthcare facilities as well as other healthcare institutions in Argentina, Uruguay, and Chile.

A major benefit of the local interface vocabulary is its size and coverage. Two previous works at HIBA are relevant to the work presented here. First, string similarity approaches [6] are used to map new description terms to existing terms in the interface vocabulary. Second, a terminology search-engine server was implemented [12]. We used a similar information retrieval approach to map a description term to the corresponding ICD-10 code. We constructed an ICD-10 search-engine that given a description term as a query, it finds the most similar existing term with an ICD-10 code and reports the corresponding ICD-10 code. We used the HIBA terminology data mapped to ICD-10 and the Spanish 2014 version of ICD-10 data. We also used HIBA terms frequency, stopwords, synonyms, collocations, abbreviations and frequent typographical errors term expansion. The ICD-10 data has 14,359 codes but HIBA mapping has 343,072 distinct terms that match only 8,537 codes (59.45%). We also added all the description terms corresponding to ICD-10 codes. We discovered a performance of 66.35% of recall @1, as an automated process [11,12]. We also evaluated recall @3, @5, @10. It should be noted that we obtained more than 80% with recall @5. These values are relevant for interactive approaches, where a health-care professional uses keywords or description terms to find the appropriate ICD-10 code for a given situation.

Methods

In this study, we constructed several indices using Elasticsearch to evaluate several alternative configurations. We used a dataset consisting of 6,787 distinct terms provided by an external institution that has been encoded by HIBA terminologists. Each electronic health record term can be codified to one or more ICD-10 codes, for example: "insuficiencia respiratoria con requerimiento alto de oxígeno" ("respiratory failure with high oxygen requirement") codes are J969 and R068.

Index Construction

The following parameters were used in each index construction with Elasticsearch:

1. Each data-set was indexed using Elasticsearch search engine using both n-gram and word analyzers.
2. For each concept in the HIBA terminology, we used its frequency to rank relevant concepts for a given mapping.
3. Mapping from HIBA terminology to ICD-10 concepts is a many-to-many mapping. We implemented these strategies using concept frequency:
   a) Maximum frequency of a HIBA concept mapped to an ICD-10 code.
   b) Minimum frequency of a HIBA concept.
   c) Average frequency of all HIBA concepts mapped to an ICD-10 code.
4. If a description term is mapped to more than one ICD-10 codes, if there is a more general term it is preferred (e.g. Z88 vs. Z889). At evaluation time, more general matches were considered correct.

5. Synonyms, collocations, abbreviations and frequent typographical errors term replacement. [6].

To construct the index, we evaluated 3 alternative data-sets:

I. The Spanish 2014 version of ICD-10 data.
II. The HIBA terminology.
III. The combination of both.

Results

We used several alternative configurations to construct the search-engine index. We used n-grams combined with word analyzers (1) and the term expansion (5) as the baseline approach. We used first only the Spanish 2014 version of ICD-10 data as the dictionary corpus (I), HIBA terminology frequency, and a general response as a valid match. This configuration resulted in 41.12 % recall @1.

Using the HIBA terminology as the dictionary corpus (II), we obtained a recall @1 of 62.84% improving the performance, without considering HIBA concept weights in ICD-10 codes. Using the two-dictionary corpus (III) provides a similar result, recall @1 62.9%.

The best result, recall @1 of 66.35, was obtained using minimum frequency strategy (3.b) stopwords and considering general codes as valid compared to a more specific one.

To be sure that the results returned by the search engine have better quality, a threshold can be used. We are considering a valid match only when the score is greater than the threshold. In other words, the search engine will not return a result under the confidence value. Using the score threshold of 6,000 we have a good balance, providing a better precision without losing too much recall. In Figure 1, we show the results of recall at more than one result for interactive choose.

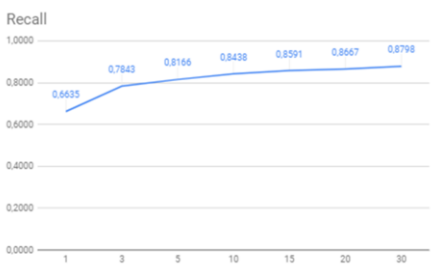

Figure 1. Recall @1, @3, @5, @10, @15, @20, @30

Discussion

We can see that the HIBA terminology to ICD-10 code mapping terms corpus, provides the best performance. The HIBA terminology covers only 59.45% of the ICD-10 codes but the richness in synonyms and frequent typographical errors enable a good result. We are using a combination of HIBA mapping to ICD-10 and the ICD-10 to complete all the ICD-10 codes but this addition doesn't provide a significant improvement.

The selection of the strategy (3.b) above, which uses the minimal frequency produces a better result than using the maximal, or the average frequency, which is somewhat surprising.

In the context of a health-care professional recording using an EHR, the system can provide the possibility of choosing the appropriate ICD-10 code from the list that returns the search engine, and as we have shown, recall @5 and @10 are very good, in the mid-80s.

Conclusions

In this work we showed several alternatives to construct an index for automatic ICD-10 term codification. We obtained a very reasonable performance. Some limitations of the current approach may be related to the differences on the corpus of the HIBA terminology data compared to the external institution data. Our approach is a dictionary projection method. A hybrid approach using supervised machine learning techniques may improve the performance.

References

[1] W.H. Organization, International statistical classification of diseases and health related problems (The) ICD-10, World Health Organization, 2004.

[2] O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM., Measuring diagnoses: ICD code accuracy, *Health services research*. **40(5p2) (**2005), 1620-39.

[3] Arifoğlu, D., Deniz, O., Aleçakır, K., & Yöndem, M., CodeMagic: semi-automatic assignment of ICD-10-AM codes to patient records, *Information Sciences and Systems* (2014), 259-268. Springer, Cham.

[4] Nitsuwat, S., & Paoin, W., Development of ICD-10-TM ontology for a semi-automated morbidity coding system in Thailand, *Methods of information in medicine* **51(06)** (2012), 519-528.

[5] Zweigenbaum, P., & Lavergne, T. (2016). Hybrid methods for ICD-10 coding of death certificates. In Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (pp. 96-105).

[6] Castaño, J., Gambarte, M. L., Park, H. J., Williams, M. D. P. A., Perez, D., Campos, F., & Zanetti, S. (2016). A machine learning approach to clinical terms normalization. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing (pp. 1-11).

[7] Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., Shaikh, K., & Al-Garadi, M. A., Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection, *PloS one* ***12*****(2)** (2017).

[8] Boytcheva, S. (2011). Automatic matching of ICD-10 codes to diagnoses in discharge letters. In Proceedings of the Second Workshop on Biomedical Natural Language Processing (pp. 11-18).

[9] Chen, Y., Lu, H., & Li, L., Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity, *PloS one* **12(3)** (2017)

[10] Jia, Z., Qin, W., Duan, H., Lv, X., & Li, H, A Hybrid Method for ICD-10 Auto-Coding of Chinese Diagnoses, *Studies in health technology and informatics*, **245** (2017) 427-431.

[11] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. New York, NY, USA: Cambridge

[12] Castaño, J., Park, H. J., Williams, M. D. P. A., Perez, D., Gambarte, M. L. Luna,D., (2018). Evaluating the Performance of a Terminology Search Engine Using Historical Data, *Medinfo* 2019.

Address for correspondence

Hee Joon Park
Ingeniería de Software
Hospital Italiano de Buenos Aires
hee.park@hospitalitaliano.org.ar