Election of Diagnosis Codes: Words as Responsible Citizens

Aron Henriksson and Martin Hassel

Department of Computer & System Sciences (DSV), Stockholm University
Forum 100, 164 40 Kista, Sweden
{aronhen,xmartin}@dsv.su.se

Abstract. Providing computer-aided support for the assignment of diagnosis codes has been approached in numerous ways, often by exploiting free-text fields in patient records. Modeling the 'meaning' of diagnosis codes through statistical data on co-occurrences of words and assigned codes—using a method known as Random Indexing—has only recently been explored as an interesting, alternative solution. It involves words in a clinician's notes 'voting' for semantically associated diagnosis codes, the election results yielding a single list of recommendations. This approach is here applied and evaluated on a corpus of over 250,000 coded patient records. The evaluation is performed by comparing the recommended codes generated by the model with those assigned by the clinicians. Applying the tf-idf weighting scheme somewhat improves results for general models (23% recall for exact matches) but has little effect on domain-specific models (32% and 59% recall for exact matches). These results confirm the potential of Random Indexing for diagnosis code assignment support, and merits further attention.

Keywords: Diagnosis Code Assignment, ICD-10, Random Indexing, Electronic Patient Records

1 Introduction

Diagnostic coding is part of a clinician's everyday work routine, its purpose being to classify diseases and other health-related issues. This makes it possible to quantify the complex operations of healthcare and thus enables effective oversight of hospitals; it also produces statistics on a regional, national and international level. For such statistics to be comparable, a standard such as the 10th revision of the *International Classification of Diseases and Related Health Problems* (ICD-10) needs to be employed [1].

The important yet somewhat tedious task of assigning appropriate diagnosis codes is typically accompanied by note-taking. This information source is often exploited in attempts to provide computer-aided coding support, given the successful application of natural language processing to other classification problems. Since a clinician's assessment and other notes often overlap with the content of the assigned diagnosis code(s), the text may be used to infer possible codes, not automating but greatly facilitating the coding process.

Computer-aided diagnostic coding support has long been an active research area (see [2] for a review); however, the results are yet to reach a level where use in a clinical setting is widespread. Natural language processing techniques are applied in most previous attempts, which include rule-based systems and statistical classifiers. Larkey and Croft [3] assign ICD-9 codes to discharge summaries using a combination of classifiers and achieve 87.9% precision, measured as the presence of the principal code among ten recommended codes.

Pakhomov et al. [4] propose a two-step classifier, where the notion of certainty is used to determine whether subsequent manual review is needed. A diagnostic statement is fed to an example-based classifier and, if unsuccessful, forwarded to a machine learning component, which generates a number of suggestions ranked by confidence. These are subjected to manual review. Fixed fields, such as gender information, are exploited to filter out improbable classifications. They report micro-averaged F_1-scores ranging from 58.6% to 96.7%, depending on whether the diagnostic entries are found in the database of previously coded entries.

In the Computational Medicine Center's 2007 Medical NLP Challenge [5], a limited set of 45 ICD-9-CM codes were to be assigned to radiology reports, more specifically based on the *clinical history* and *impression* fields. The automatic systems were evaluated against a gold standard, which was created using the majority annotation from three independent annotation sets. Many of the best-performing contributions are heavily dependent on hand-crafted rules, with the winning contribution achieving a micro-averaged F_1-score of 89%. Farkas and Szarvas [6] decided instead to combine predefined rules based on the ICD-9-CM coding guide with automated procedures, which effectively reduces the development effort and yet results in a high micro-averaged F_1-score of 88.9%. The third-best system uses classifiers that perform a binary classification for each label and achieves a micro-averaged F_1-score of 87.7%. They found that the choice of classifier was less important but that identifying negations, making use of the structure of UMLS[1] and enriching the to-be-classified documents with hypernyms helped [7].

The possible application of the *word space model* to this problem has only recently been investigated. We have previously proposed the use of *Random Indexing* as an interesting alternative [8]. In that study, the method is evaluated qualitatively on a limited set of documents, yielding promising yet tentative results. There is thus a need for those results to be consolidated by a more quantitative evaluation, while there is also considerable scope for the method to be developed further.

The word space model is an application of the vector space model (see [9] for a review) and attempts to capture the meaning of words through statistics on word co-occurrences. This is based on the *distributional hypothesis*, which states that words that appear in similar contexts tend to have similar properties. Thus if words repeatedly co-occur, we can assume that they in some way refer to similar concepts [10]. Given the successful application of word space models to information retrieval, semantic knowledge tests (e.g. TOEFL), text categoriza-

[1] Unified Medical Language System

tion, text summarization, word sense disambiguation, etc., it may also prove a viable solution for diagnostic coding support.

2 Method

Random Indexing [11], [12] is applied on a corpus of almost 270,000 coded patient records to calculate co-occurrences of words in clinical notes and assigned ICD-10 codes, creating models which can be used to predict diagnosis codes for uncoded documents. The method comprises the following steps: (1) pre-processing the data, (2) building a number of word space models using the training data, (3) generating suggested diagnosis codes for the documents in the test data and (4) evaluating the results by matching the suggested codes with those assigned by the clinicians.

A subset of the *Stockholm EPR* corpus [13] is used for training and evaluating the created models. The subset contains approximately 5.5 million notes from 838 clinical units[2]. Documents contain notes from one clinical unit about a single patient and are created differently depending on the definition of a *patient visit*: (1) notes made on the same day, (2) notes made on consecutive days and (3) notes made with at most one undocumented day in between. The reason for the different partitions is to see what effect the length of the documents has on the models: do they benefit from additional information or are they impaired by potentially less uniform content? Documents are associated with assigned codes; documents without any associated codes are ignored. These documents are first pre-processed: lemmatization is done using the *Granska Tagger* [14], while punctuation, digits and stop words are removed. The pre-processed data is subsequently partitioned into two subsets, where one is used for training (90%) and the other is set aside for testing (10%). In the training set, the associated ICD-10 codes are included in the input documents, whereas in the test set, they are retained separately for evaluation. Diagnosis codes are generally not mentioned in the free-text fields; the idea is that no such information should be revealed to the models in the testing phase.

We then generate a number of models by constructing a word vector for each token (words and ICD-10 codes) encountered in the training data. The relative directions of these word vectors in the word space are assumed to indicate semantic similarity. The construction of the word vectors requires the context to be defined and represented in some way. The context is often defined as a paragraph or a window of surrounding words; however, in our case, it includes an entire document, as there is no sequential dependency between the diagnosis code and the words in the document. Each document, i.e. the context, is assigned a unique and randomly generated context vector, which is a sparse, high-dimensional and ternary vector. The dimensionality of the context vectors depends on the size and redundancy of the data—we have set it to 1,000—with a very small number (1-2%) of randomly distributed +1s and -1s, while the remaining elements are set

[2] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5.

to 0. The word vectors are then built by processing the documents: every time a word occurs in a particular context, the context vector is added to the word vector. The usage of a word is thus represented as a vector, which is the sum of the context vectors of the contexts in which it appears. We build a model for each of our data sets: three general models (*general elections*), which are trained on the entire data set, and two domain-specific models (*municipal elections*), which are trained on subsets from a single type of clinic: ENT (Ear-Nose-Throat) and rheumatology clinics respectively. The reason for creating domain-specific models is that, by limiting the sparsity of the data and the classification problem (# of possible ICD-10 codes), we may achieve better results. Furthermore, we build bigram versions of some of the above models. As bigrams are more informative than unigrams and represent a means of dealing with multiword expressions, they may prove to be appropriate units in the construction of the models.

These models are then used to produce a ranked list of a number of recommended ICD-10 codes (e.g. 10) for each query document (i.e. excluding codes). For each word in a document, a ranked list of semantically correlated words is retrieved from the model. As we are interested in recommending diagnosis codes, the results are filtered to include only such tokens. The individual lists of all the words in the document are combined to produce a single ranked list. This *ensemble method* is carried out in one of two ways: (1) by using the ranking positions of the codes in the individual lists and (2) by using the semantic (cosine) similarity scores of the codes in the individual lists.

In this initial approach, we apply the 'one word, one vote' system, i.e. all words have an equal say in electing the diagnosis codes (*democratic approach*). This simple yet rather naive approach is refined by implementing the tf-idf (term frequency-inverse document frequency) weighting scheme (*meritocratic approach*), effectively giving a stronger voice to prominent (tf) words that have a high discriminatory value (idf). Tf is implicit in our method, as each *instance* of a word has its own vote, while idf is retrieved from the model.

The evaluation is conducted by comparing the codes that were assigned by the clinicians with the model-generated recommendations. This matching is done on all four possible levels of ICD-10 according to specificity (Figure 1). If a clinically assigned code is not given at the most specific level—as is often the case—and that same code is recommended by the model, it naturally counts as an exact match. To gauge the quality of the results, they are compared with a naive baseline for each model. It is created by matching the assigned codes for each document against a list of the most frequent labels in the training set.

Fig. 1. *The structure of ICD-10 allows division into four levels.*

3 Results

The data sets have a distinct number of tokens and unique codes per visit (Table 1). While the set of codes is identical across the general data sets (12,396 labels), these are much smaller in the domain-specific versions (1,713 and 638 labels).

Table 1. *Data set statistics.*

data set	**visits**	**codes**	**codes/visit (max)**	**tokens/visit (min-max)**
0days_general	∼278 k	12,396	1.7 (47)	165.6 (1-3913)
1days_general	∼274 k	12,396	1.7 (47)	177.7 (1-3913)
2days_general	∼271 k	12,396	1.7 (47)	186.1 (1-3913)
0days_ENT	∼24 k	1,713	2.1 (20)	138.2 (1-621)
0days_Rheumatology	∼9 k	638	1.2 (16)	88.5 (1-549)

The models trained and evaluated on the bigram versions of the above data sets fail to yield an improvement over the simpler unigram models. Those models are therefore not considered in any of the subsequent experiments. Out of the two ensemble methods applied in the production of the final list of recommended codes, the one whereby the cosine similarity scores are taken into account generally performs somewhat better than the more basic variant in which the ranking positions are used.

The models are initially employed in a democratic approach, i.e. all words have an equal vote (Table 2). In the general data sets, approximately 21% of the assigned codes are recommended by the corresponding models. Partial matches are more frequent, with approximately 29% matched on level 2. The results are higher when applying the domain-specific models: 31% exact matches in ENT and 59% in Rheumatology. Similar increases are observed when matching at the less specific levels, with 61% and 93% partial matches respectively.

Table 2. Model.

Model	Recall (top 10)			
	E	L3	L2	L1
0days_general	0.21	0.23	0.29	0.59
1days_general	0.23	0.25	0.33	0.60
2days_general	0.20	0.22	0.29	0.59
0days_ENT	0.31	0.31	0.39	0.61
0days_Rheumatology	0.59	0.59	0.70	0.93

The models are then applied in a meritocratic fashion, i.e. some words have a greater say than others (Table 3). This approach leads to a small increase in the performance of the general models (up to 3 percentage points), with little impact on the domain-specific models, even having an adverse effect in Rheumatology.

Table 3. *Meritocratic approach* – general and municipal elections. Overall recall (top 10) for all four possible levels of ICD-10.

Model	Recall (top 10)			
	E	L3	L2	L1
0days_general	0.23	0.25	0.32	0.59
1days_general	0.23	0.25	0.33	0.60
2days_general	0.22	0.24	0.32	0.60
0days_ENT	0.32	0.33	0.39	0.61
0days_Rheumatology	0.59	0.59	0.69	0.93

When we increase the number of generated recommendations, the results improve to some degree (Figure 2). For recall 15, the results increase by up to 7 percentage points and, for recall 20, they increase by up to 13 percentage points. That is, when 20 recommended codes are generated, 28% of the clinically assigned codes are recommended by the general models, while 38% (ENT) and 69% (Rheumatology) are matched by the domain-specific models. The baseline is beaten by the general models but not always by the domain-specific versions.

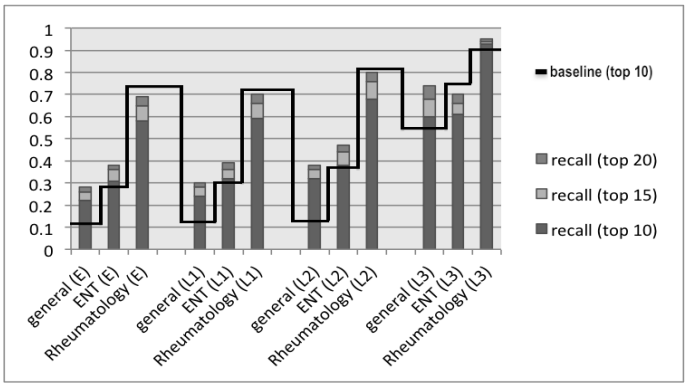

Fig. 2. Improvements observed when increasing the number of recommendations from 10 to 15 and 20. The baseline is for recall top 10 only.

4 Discussion

The definition of a patient visit, which potentially determines the length of the documents and the number of assigned codes, appears to have a negligible impact on the results. As can be seen in Table 1, the average number of tokens per visit only increases slightly, while the average number of assigned codes is more or

less equal. The reason for this is probably that the status of inpatients is usually monitored and documented on a near-daily basis.

The domain-specific models perform significantly better than the general models. This was expected considering the extent to which the classification problem is thereby limited: the number of labels is reduced from 12,396 to 1,713 (ENT) and 638 (Rheumatology). The discrepancy between the two domain-specific models, including the average number of codes per visit in ENT (2.1) and Rheumatology (1.2), is likewise a possible reason for the latter performing better than the former.

The meritocratic approach had surprisingly little effect on the results. While it did have a positive yet small impact on the general models, it had little—and even a negative—impact on the domain-specific models. A qualitative analysis of which types of words benefit from this weighting scheme would be interesting. It is likely that a small set of keywords are highly indicative of the diagnoses and need to be given additional weight. In a *technocratic approach*, this could be achieved by seeking out words used in the ICD-10 descriptions or by looking for SNOMED CT terms, such as diseases and body parts.

The fact that the bigram models performed worse than the unigram models could possibly be due to the sparsity of the data. That is, there are many more distinct tokens and fewer instances of each. The same conclusion was reached by Suominen et al. [7], where bigrams and trigrams were shown to yield no improvement in the assignment of ICD-9-CM codes.

The results presented here are not directly comparable with those in previous studies, primarily due to the tasks being of different orders of magnitude. In comparison to the 2007 shared task [5], in which a limited set of 45 labels were used, our classification problem comprises thousands of labels. While they ensured that the testing data did not contain any unseen labels, ours could and did. Moreover, the shared task was limited to assigning one or two codes, although they were punished for failing to assign the exact the number of labels. In contrast, we have so far only presented recall scores, measured as the presence of clinically assigned codes in a list of 10, 15 or 20 recommendations. The same is done in [3], although, in that case, only a single label—the principal code—was to be assigned. In contrast to the shared task, where the data was produced by expert coders and the best-performing systems to a large extent relied on handcrafted rules, this study evaluates a statistically-based method on large volumes of clinically generated data. However, building models on real, noisy data, without relying heavily on rules that are expensive to create, is precisely what could make this method a feasible solution for future clinical coding support.

In future work, we plan to exploit fixed fields in patient records, such as age and gender, to avoid statistically rare correlations, as was successfully done in [4]. Furthermore, we believe negation handling may have a positive effect on results. This could be achieved by, for instance, ignoring negated diagnoses in the construction of the word space models. Different levels of certainty may also be factored into the equation. Detection of negation and uncertainty was shown in [7] to have a positive effect on the automatic assignment of diagnosis codes.

5 Conclusion

We have quantitatively evaluated the use of Random Indexing as a means to provide diagnostic coding support on over 250,000 patient records. An array of models and two ensemble methods were evaluated. A meritocratic approach (tf-idf weighting) yields little improvement over a democratic approach (one word, one vote) in the election of appropriate diagnosis codes. Domain-specific models produce significantly better results (at best 61% exact matches and 93% partial matches) than general models (22% exact matches and 61% partial matches).

References

1. World Health Organization: International Classification of Diseases (ICD). [Internet]. Geneva: WHO; 2011 [accessed June 2011, available from: `http://www.who.int/classifications/icd/en/`] (2011).
2. Stanfill, M.H., Williams, M., Fenton, S., Jenders, R.A., Hersh, W.R.: A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc, 17, pp. 646–651 (2010).
3. Larkey, L.S., Croft, W.B.: Automatic Assignment of ICD9 Codes to Discharge Summaries. In PhD thesis University of Massachusetts at Amherst, Amherst, MA (1995).
4. Pakhomov, S.V.S., Buntrock, J.D., Chute, C.G.: Automating the Assignment of Diagnosis Codes to Patient Encounters Using Example-based and Machine Learning Techniques. J Am Med Inform Assoc, 13, pp. 516-525 (2006).
5. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermal, D.J., Johnson, N., Cohen, K.B., Duch, W.: A Shared Task Involving Multi-label Classification of Clinical Free Text. In Proceedings of the Workshop on BioNLP (2007).
6. Farkas, R., Szarvas, G.: Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics, 9(3) (2008).
7. Suominen, H., Ginter, F., Pyssalo, S., Airola, A., Pahikkala, T., Salanter, S., Salakoski, T.: Machine Learning to Automate the Assignment of Diagnosis Codes to Free-Text Radiology Reports: A Method Description. In Proceedings of the IMCL/UAI/COLT Workshop on Machine Learning for Health-Care Applications (2008).
8. Henriksson, A., Hassel, M., Kvist, M.: Diagnosis Code Assignment Support Using Random Indexing of Patient Records — A Qualitative Feasibility Study. In Proceedings of AIME, 13th Conference on Artificial Intelligence in Medicine (2011).
9. Turney, P. D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, pp. 141–188 (2010).
10. Harris, Z. S.: Distributional structure. Word, 10, pp. 146–162 (1954).
11. Sahlgren, M: Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels. In Proceedings of Semantic Knowledge Acquisition and Categorization Workshop at ESSLLI'01 (2001).
12. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. In PhD thesis Stockholm University, Stockholm, Sweden (2006).
13. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249 (2009).
14. Knutsson, O., Bigert, J., Kann, V.: A Robust Shallow Parser for Swedish. In Proceedings of Nodalida (2003).