Diagnosis Code Assignment Support Using Random Indexing of Patient Records – A Qualitative Feasibility Study

Aron Henriksson[1], Martin Hassel[1], and Maria Kvist[1,2]

[1]Department of Computer and System Sciences (DSV), Stockholm University
Forum 100, 164 40 Kista, Sweden
[2]Department of Clinical Immunology and Transfusion Medicine, Karolinska University Hospital, 171 76 Stockholm, Sweden

Abstract. The prediction of diagnosis codes is typically based on free-text entries in clinical documents. Previous attempts to tackle this problem range from strictly rule-based systems to utilizing various classification algorithms, resulting in varying degrees of success. A novel approach is to build a word space model based on a corpus of coded patient records, associating co-occurrences of words and ICD-10 codes. Random Indexing is a computationally efficient implementation of the word space model and may prove an effective means of providing support for the assignment of diagnosis codes. The method is here qualitatively evaluated for its feasibility by a physician on clinical records from two Swedish clinics. The assigned codes were in this initial experiment found among the top 10 generated suggestions in 20% of the cases, but a partial match in 77% demonstrates the potential of the method.

Keywords: ICD-10 Assignment, Random Indexing, Electronic Patient Records, Qualitative Evaluation

1 Introduction

1.1 Diagnosis code assignment support

Patient records comprise a combination of structured information and free-text fields. Free-text fields allow healthcare personnel to record observations and to reason about possible diagnoses and actions in a flexible manner. Fixed fields, i.e. closed classes, are on the other hand often desirable as they can easily be aggregated to produce meaningful statistics. The 10th revision of the *International Classification of Diseases and Related Health Problems* (ICD-10) is a classification system that is used to record medical activity. Its main purpose is to enable classification and quantification of diseases and other health-related issues [1].

Assigning ICD-10 codes is a necessary yet time-consuming task that keeps healthcare personnel away from their core responsibility: tending to patients. To facilitate the selection of diagnosis codes among a myriad of options, computer-aided coding support has been an active research area for the past twenty years or so (see [2] for a literature review).

The most common approach is to base coding support on natural language processing of clinical documents. This study is to all intents and purposes similar to that of Larkey and Croft [3]. They assign ICD-9 codes to discharge summaries using three classifiers trained on a prelabeled corpus. By giving extra weight to words, phrases and structures that provide the most diagnostic evidence, results are shown to improve. A combination of classifiers yields a precision of 87.9%, where the principal code is included in a list of ten recommendations.

A large number of related studies were sparked by the Computational Medicine Center's 2007 Medical NLP Challenge[1], where a limited set of 45 ICD-9-CM codes were to be assigned to free-text radiology reports. Many of the solutions are hand-crafted rule-based systems, giving at best an 89.1% average F-score.

1.2 Word space models

A common trait of the aforementioned methods is that they in some way attempt to represent features of the classified text passage. Word space models constitute a family of models that capture meaning through statistics on word co-occurrences. Since its introduction, *Latent Semantic Analysis* (LSA) [4] has more or less spawned an entire research field with a wide range of word space models as a result. Numerous publications report exceptional results in many different applications, such as information retrieval, various semantic knowledge tests (e.g. TOEFL), text categorization and word sense disambiguation.

The idea behind word space models is to use statistics on word distributions in order to generate a high-dimensional vector space. Words are here represented by context vectors whose relative directions are assumed to indicate semantic similarity. The basis of this is the *distributional hypothesis*, according to which words that occur in similar contexts tend to have similar properties (meanings/functions). If we repeatedly observe two words in the same contexts, it is not too far-fetched to assume that they also refer to similar concepts [5].

2 Method

We employ the *Random Indexing* word space approach [5], [6], which presents an efficient, scalable and inherently incremental alternative to LSA-like word space models. The construction of context vectors using Random Indexing can be viewed as a two-step process.

First, each context—defined as the document or paragraph in which a word occurs, or as a (sliding) window of a number of surrounding words—is assigned a unique and (usually) randomly generated label. These labels are sparse, high-dimensional and ternary vectors. Their dimensionality d is usually in the range of a couple of hundred to several thousands, depending on the size and redundancy of the data, and they consist of a very small number (usually about 1-2%) of randomly distributed +1s and -1s, with the rest of the elements in the vectors set

[1] `http://www.computationalmedicine.org/challenge/2007/index.php`

to 0. Next, the actual context vectors are produced by scanning the text: each time a token w occurs in a particular context, the d-dimensional random label of that context is added to the context vector of w. Thus, each context that a token w appears in has an effect, through its random label, on the context vector of w. Words are in this way effectively represented by d-dimensional context vectors, which are the sum of the random labels of the co-occurring contexts.

In our experiments, we set the context to an entire document, as there exists no sequential dependency between the diagnosis code and the words. A document contains all free-text entries concerning an individual patient made on consecutive days at a single clinic, as well as the associated ICD-10 codes. The data used for training and testing the model is a part of the *Stockholm EPR* corpus [7] and contains 273,888 documents written in Swedish, 12,396 distinct ICD-10 codes and 838 clinical units[2]. A document contains an average of 96 words and 1.7 ICD-10 codes. All documents are first pre-processed. In addition to lemmatization, which is done using the *Granska Tagger* [8], punctuation, digits and stop words[3] are removed. The data is then split 90:10 between training and testing, where the training documents include the associated codes, while, in the testing data, they are retained separately for evaluation.

In the testing phase, a document is input: for each word, a ranked list of semantically correlated words is produced by the model. As our interest lies in assigning ICD-10 codes only, the result set is restricted to such tokens. The lists for each of the words in the document are combined to yield a single ranked list of ten ICD-10 codes. The results are manually evaluated by a physician on a total of 30 documents: 15 from an emergency ward and 15 from a rheumatology clinic. The recommended codes are evaluated for their relevance compared to the clinical text and matched with the codes assigned by the physicians.

3 Evaluation

We found the assigned code among the ten suggestions in 20% of the cases (Table 1). It should be noted that, although this number seems low, matching the assigned code exactly is difficult, given the structure of ICD-10. The codes contain four levels, going from general categories to more specific descriptions, with codes typically assigned at the two most specific levels. As a result, the model will sometimes generate suggestions that are either more or less specific than the assigned code(s), even if they are otherwise similar. For partial matches[4], however, there was at least one match in 77% of the cases. Another complicating factor is that some closely related diagnoses belong to different categories altogether, resulting in a reasonable suggestion not even yielding a partial match.

We therefore also evaluated to what extent the recommended codes were reasonable. On average, 23% of the suggestions were reasonable in the sense

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

[3] Frequently occurring words that have a low discriminatory value.

[4] A partial match is defined as matching the assigned code on any of the levels.

that they were deemed possible diagnoses by a physician based on the clinical text. Some of these—with occurrences in more than half of the documents—were diagnoses that were not assigned a code, either because they could not yet be confirmed or due to their lesser degree of relevance in relation to the problem at hand. We also measured the proportion of the suggested diagnoses which had a palpable connection—in the form of clinically significant words—to the text. We found such a word-based connection in 45% of the recommended codes.

Table 1. Evaluation of recommended codes in relation to input document and assigned code. Word connections and reasonable suggestions are calculated for all recommended codes (10/doc), while full and partial matches are calculated on a document level.

	Word connection	Reasonable suggestion	Assigned code in top 10	
			Full match	Partial match
Rheumatology	43% (±25)	23% (±18)	33%	87%
Emergency	47% (±16)	23% (±13)	6.7%	67%
Overall	45% (±21)	23% (±16)	20%	77%

The results between the two clinics are fairly similar when it comes to word connections and reasonable suggestions, which indicates that the method is equally applicable to two very different types of clinics. The difficulty of the task is not necessarily identical, however, as the results of the full and partial matches demonstrate. A possible explanation for this could be that a rheumatology clinic consists of a more homogenous group of specialists using a limited set of well-known diagnosis codes, whereas the number of possible codes is likely to be larger in an emergency ward, where the specificity of the code may also depend on the specialty of the coder.

4 Discussion

Even if ICD-10 codes may sometimes be used in a complementary manner to the free-text fields in patient records, in our evaluation there was almost always a connection between the text and the assigned code(s). This is encouraging, as it constitutes a prerequisite for the application of Random Indexing—or any other implementation of the word space model—to be successful in recommending diagnosis codes. There are, however, a number of limitations to the method, especially in its current mold.

In addition to the difficulties mentioned earlier, a big challenge is posed by the present inability to capture the function of negations. This is particularly problematic when applying the method to the clinical domain, where ruling out possible diseases, symptoms and findings is inherent in the operations of clinical practice. The consequence is, of course, that diagnosis codes will be associated with symptoms and findings that are negated in the text.

In this rather naive implementation, all words have an equal say in "voting" for the most appropriate diagnosis code. A means of countering this is to incorporate some form of weighting. We plan initially to employ the well-established

tf-idf (term frequency-inverse document frequency) weighting scheme, which is based on the prominence of tokens (*tf*) and their discriminatory value (*idf*). Giving more weight to words in the ICD-10 descriptions and to certain sections of the patient record may also yield improved results.

Given the large amount of possible diagnosis codes, domain-specific models trained on a single type of clinic may possibly perform better. Taking advantage of structured information, such as age and gender, is also likely to limit the classification problem by ruling out, or giving less weight to, unlikely correlations. Finally, building word space models based on bigrams, which have a higher discriminatory value than unigrams, is also something we plan to investigate.

Once we have implemented some of these features, we will also conduct a quantitative study, in which all the test data will be evaluated in relation to the assigned diagnosis codes.

5 Conclusion

We have introduced Random Indexing of patient records as a new approach to the problem of diagnosis code assignment support. The outcome of the qualitative evaluation is fairly encouraging—particularly as it is here applied in a rather elementary fashion—with 20% full matches and 77% partial matches against a list of ten recommended ICD-10 codes. With further room for improvement, it may prove an efficient and effective solution.

References

1. World Health Organization: International Classification of Diseases (ICD). [Internet]. Geneva: WHO; 2010 [accessed February 2010, available from: `http://www.who.int/classifications/icd/en/`] (2010).
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. Landauer, T.K., Foltz, W., Laham, D.: Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259–284 (1998).
5. 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).
6. 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).
7. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249 (2009).
8. Knutsson, O., Bigert, J., Kann, V.: A Robust Shallow Parser for Swedish. In Proceedings of Nodalida (2003).