Assigning clinical codes with data-driven concept representation on Dutch clinical free text

Elyne Scheurwegs[a,b,*], Kim Luyckx[c], Léon Luyten[d], Bart Goethals[a], Walter Daelemans[b]

[a] *University of Antwerp, Advanced Database Research and Modelling research group (ADReM) Middelheimlaan 1, B-2020 Antwerp, Belgium*
[b] *University of Antwerp, Computational Linguistics and Psycholinguistics (CLiPS) research center Lange Winkelstraat 40-42, B-2000 Antwerp, Belgium*
[c] *Antwerp University Hospital, ICT department Wilrijkstraat 10, B-2650 Edegem, Belgium*
[d] *Antwerp University Hospital, Medical information department Wilrijkstraat 10, B-2650 Edegem, Belgium*

Abstract

Clinical codes are used for public reporting purposes, are fundamental to determining public financing for hospitals, and form the basis for reimbursement claims to insurance providers. They are assigned to a patient stay to reflect the diagnosis and performed procedures during that stay. This paper aims to enrich algorithms for automated clinical coding by taking a data-driven approach and by using unsupervised and semi-supervised techniques for the extraction of multi-word expressions that convey a generalisable medical meaning (referred to as *concepts*). Several methods for extracting concepts from text are compared, two of which are constructed from a large unannotated corpus of clinical free text. A distributional semantic model (i.c. the word2vec skip-gram model) is used to generalize over concepts and retrieve relations between them. These methods are validated on three sets of patient stay data, in the disease areas of urology, cardiology, and gastroenterology. The datasets are in Dutch, which introduces a limitation on available concept definitions from expert-based ontologies (e.g. UMLS). The results show that when expert-based knowledge in ontologies is unavailable, concepts derived from raw clinical texts are a reliable alternative. Both concepts derived from raw clinical texts perform and concepts derived from expert-created dictionaries outperform a bag-of-words approach in clinical code assignment. Adding features based on tokens that appear in a semantically similar context has a positive influence for predicting diagnostic codes. Furthermore, the experiments indicate that a distributional semantics model can find relations between semantically related concepts in texts but also introduces erroneous and redundant relations, which can undermine clinical coding performance.

Keywords: Clinical Coding, Data Mining, Text Mining, Unsupervised learning, International Classification of Diseases, Electronic Health Records, Distributional Semantics, Word2Vec

1. Introduction

Medical knowledge is electronically stored in a high number of complex data sources, such as electronic health records (EHRs), electronic archives, ontologies, and scientific publications [1, 2, 3]. In a modern hospital setting, clinical codes are determined based on information found in the electronic health record. These clinical codes are assigned primarily for the purpose of reporting and reimbursement from health care providers or governments. Their widespread adoption in clinical environments allows for the usage as an important and complementary factor in research applications (e.g. identifying acute venous thromboembolisms) [4]. While clinical codes are often assigned manually by a team of specialized coders, techniques that can (semi-)automatically predict these codes can lower the burden of this codification process.

Most data stored in hospitals is not annotated due to the large effort that is required from physicians to accurately annotate this data. This limits the usability of this data for supervised machine learning techniques. We investigate unsupervised and semi-supervised techniques to create appropriate text representations for use in a prediction pipeline for clinical codes. The objective of this paper is to improve automated prediction of clinical codes by (I) introducing methods that are independent of expert-created ontologies to extract these concepts from the source documents of this patient stay and (II) using a distributional semantics model to generalize and represent concepts associated with a patient stay.

*Corresponding author
Email address: elyne.scheurwegs@uantwerpen.be (Elyne Scheurwegs)

1.1. Background

Adequate feature engineering is arguably one of the most important steps for any sort of machine learning task, including trying to learn from unstructured clinical documents. The feature engineering task here can be defined as the conversion of unstructured data into a structured representation suitable to make predictions. A simple strategy is to use a lexical representation by using a library of relevant tokens (words occurring in a medical dictionary), or just using the text itself as the library (e.g. bag-of-words in which all unique tokens occurring in the document are counted and directly used as a representation). Other strategies include using a syntactic representation or a semantic representation. A lexical representation can be enhanced (or filtered) with semantic and/or syntactic properties as metadata (e.g. PoS-tags).

In this work, a series of documents, associated with a patient stay, is represented by a series of extracted concepts. These concepts are in essence multi-word expressions that convey a generalisable medical meaning.

1.1.1. Medical Information Extraction

The approach chosen to extract information from clinical free texts is largely determined by the intended purpose. This purpose can be a specific case (e.g. identifying heart failure diagnostic criteria [5]), or a generic task (e.g. extracting medication terms from clinical narratives) [6]. In the ShARe/CLEF 2013 eHealth shared task [7], entities were recognized in clinical notes and subsequently normalized to UMLS identifiers (i.e. CUI codes) [8]. The best-ranking system found entities with supervised machine learning techniques, for which candidate CUIs were represented as Bag-of-Words, weighted with their TF-IDF score [9, 10].

Pathak et al. mapped structured and unstructured data onto the UMLS identifier structure for the purpose of high-throughput phenotyping [11]. This approach allows for the integration of multiple types of data sources, but is substantially dependent on the existence of predefined expert knowledge in ontologies and vocabularies. This is particularly problematic for languages with a smaller number of medical lexicons, such as Dutch.

1.1.2. Distributional semantics in medical corpora

A distributional semantic model (DSM) acquires a semantic representation for tokens by looking at the surrounding tokens in a large corpus [12]. Tokens are assumed to be semantically related if they are often surrounded by similar context. Antonyms and frequently co-occurring tokens are thus also marked as semantically related (e.g. 'white' and 'black','dear' and 'colleague' in headings). Jonnalagadda et al. extracted medical information from clinical narratives with a Random Indexing (RI) DSM [13, 14]. They retrieved semantically related tokens in an unannotated corpus of Medline abstracts with the RI model, after which they supplemented the basic features (i.e. dictionary- and pattern-matched features and Part-of-Speech tags) in a machine learning algorithm with the related tokens. Including semantically related tokens increased their achieved F-measure to 91.3% for inexact matches (an increase of 2%). Henriksson et al. similarly applied RI to enhance a medical lexicon with synonyms and abbreviations [15].

Moen et al. applied two distributional semantic models (Random Indexing and a word2vec model [16]) to retrieve care episodes that are similar to the care episode under review [17]. A care episode consisted of a free text summary. Their most successful variant modified the network creation of the word2vec skip-gram model by introducing feedback that takes the ICD-10 code assigned to the training samples into account [18]. While this method significantly improved results, it also required an ICD-code to be linked to each document used to train the word2vec method. This is often not the case with archived documents. The second best variant was the unmodified word2vec skip-gram model.

In this study, we chose to use the word2vec skip-gram model [16, 19]. Word2vec is an implementation of two vector representation algorithms (CBOW and skip-gram) for tokens. These algorithms both encompass a neural network, consisting of one input layer, one hidden layer, and one output layer. The vocabulary items are mapped to each input node, and a hidden layer within the model is shaped with n nodes (with n representing the number of dimensions requested), with each node representing one dimension of the desired vector. The models are then trained by presenting them with each vocabulary item and the context in which it occurs. This process is repeated until the network converges to a predetermined output error. A trained model provides a multi-dimensional space in which each word and/or token is represented by an individual dense vector with a relatively low number of dimensions.

1.1.3. Automated Clinical Coding

Current automated clinical coding approaches are often used in controlled environments, with strongly normalized data and a limited scope in document type (e.g. radiology reports) and disease area (e.g. oncology) [20]. The most successful approaches are (partially) handcrafted, which renders them harder to port to different languages or medical specialties [21]. Perotte et al. predicted 5030 unique ICD-9-CM codes on discharge files from the MIMIC-II dataset by exploiting the ICD-9-CM hierarchy, with a resulting F-measure of 39% [22, 23]. Scheurwegs et al. integrated structured and unstructured data sources to assign clinical codes to patient stays for multiple specialties [24]. They confirmed the large difference in

achieved F-measure between specialties and presented a technique that is portable over medical specialties. The texts in the discharge files of the latter approaches was represented with a Bag of Words (BoW) approach, rendering the approach more portable over different languages.

This paper aims to show the feasibility of using unsupervised methods for representing unstructured data in automated clinical coding approaches. Unsupervised methods are used to both detect concepts in texts and represent those concepts in a dense vector space. The proposed methods are mainly dependent on unannotated resources (raw text) instead of on annotated resources (such as ontologies, hand-crafted rules for information extraction, and annotated training data) and are thus easily deployable on languages with limited coverage in ontologies. These methods are evaluated on a medical dataset in Dutch.

2. Materials and methods

2.1. Dataset

Two types of datasets are derived from the clinical data warehouse at the Antwerp University Hospital. The first dataset consists of an unannotated corpus of 2,374,723 automatically de-identified texts (with an average of 152 words per text). This data in this corpus is essentially raw text and covers multiple medical specialties. The second dataset consists of a randomized subset of anonymized patient stays with associated documents (radiology reports, requests, surgery reports, notes, letters, and attestations) and ICD-9-CM codes [18]. This dataset is divided into three specialties (i.e. cardiology, gastroenterology, and urology, with respectively 10000, 7440, and 3440 patient stays). In table 1, we show the total number of texts, patient stays and the properties of both diagnostic and procedural codes in each dataset.

The task is defined as predicting all clinical codes (i.e. procedural codes, primary as well as secondary diagnosis codes) associated with a patient stay, given all associated clinical documents. An estimate of the relative frequency of different text categories is seen in figure 1. Patient records are anonymized, but not filtered in any way. Duplicates of documents might be present, as well as patient stays that do not contain any documents.

2.2. Data preprocessing

Both datasets mentioned above are preprocessed with several low-level natural language processing (NLP) steps (sentence splitting, tokenization, lemmatization, part-of-speech tagging, and chunk tagging). Frog, a morpho-syntactic analyzer and dependency parser for Dutch text, fulfils this task [25]. The output of these modules is used in the approaches presented in the following sections.

2.3. Concept detection

A Bag-of-Words (BoW) representation of a text is commonly used in a text categorization task using machine learning algorithms. This method essentially splits up the text into a list of words and uses the absolute occurrence of each word as a separate feature. This approach is relatively robust, and language-independent. However, a lot of information is lost when a text is converted to a BoW, such as relations between words (e.g. high fever), the context in which a word occurs (e.g. no sign of hypothermia), and items that span multiple words (e.g. Diabetes Mellitus). While including bigrams and trigrams can solve the problem of capturing items spanning multiple words, this would also dramatically increase the number of features, shifting the problem to a feature selection and noise-reduction problem.

The aforementioned lost information can be retained if the text is converted to a list of concepts of one or more words, with metadata containing information about the context. As we are only interested in retaining medical facts or hypotheses occurring in text, this type of information extraction will suit our needs. Several methods of information extraction have already been introduced, but we focus on methods that are minimally dependent on annotated training data while still performing well on this specific task. More specifically, we used a linguistic pattern extraction method based on pointwise mutual information (which we will refer to as linguistic mutual information or LMI), and a bootstrapped pattern mining method (BPM), as introduced by Gupta and Manning [26]. We supplement this with a dictionary-based approach (DICT), for additional performance and for comparison purposes.

2.3.1. Dictionary-based approach (DICT)

For the dictionary-based approach (DICT), several lexicons were combined: 3BT (a Belgian bilingual corpus of Dutch and French medical terms), Dutch terms from the UMLS metathesaurus [8], and a list of European brand names and active compounds of medication [27]. Each word in the text is retrieved in these lexica. When a match is found, we extract the surrounding base phrase chunk (parts of a sentence, such as a prepositional phrase and noun phrases). By extracting the base phrase chunk as opposed to taking a window, we avoid trailing verbs and adjectives that are not linked to the concept. Then each retained phrase chunk is compared with the raw definition of medical terms in the UMLS, 3BT and medication lexicons. If all words in the definition are found within the word chunk, the concept is added to the concept list. This is considered a rule-based approach. As seen in figure 2A, this method is prone to missing concepts due to the limited coverage of the lexicon. For dictionary-based concepts, we found that 47% of the candidate concepts are retained as a concept, with an

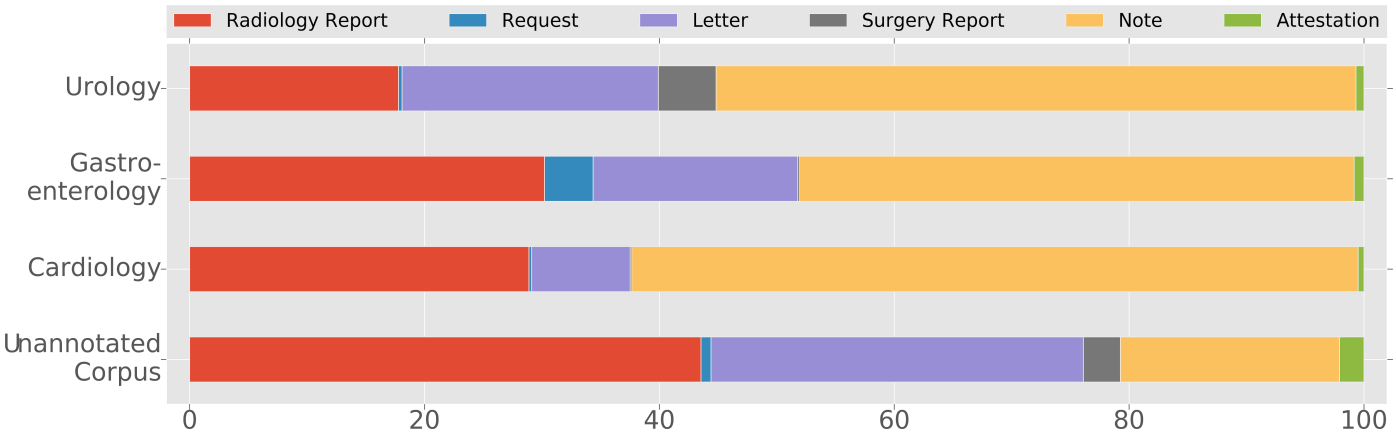

Figure 1: Ratios of the different document types (in %) for the unannotated dataset and for each disease area.

Table 1: Dataset description, per specialty. Label cardinality indicates the average number of assigned labels per sample.

	Cardiology	Gastrology	Urology
Total patient stays	10k	7.44k	3.44k
Total texts	90.5k	86.4k	36.9k
Label cardinality for diagnosis codes	7.9	4.87	5.83
N labels for diagnosis codes	2152	2165	1422
Label cardinality for procedure codes	3.04	0.87	1.09
N labels for procedure codes	230	232	282

average of 16 concepts per text. A dictionary-based concept is represented by its identifier. In the case of UMLS, each concept is normalised to the underlying CUI.

2.3.2. Linguistic PoS and pointwise mutual information-based approach (LMI)

The linguistic PoS and pointwise mutual information method (linguistic mutual information or LMI) is illustrated in figure 2B. This unsupervised method uses predefined sequences of PoS-tags (e.g. Adjective-Noun, Noun-Noun; see table 3 in supplementary) to convert the raw text into a list of candidate concepts. These candidates are then scored based on their relative frequency of occurrence of each word in the candidate concepts in both a medical corpus (consisting of the historical dataset) and in multiple general corpora (i.e. SoNaR subtypes [28]). The different types of text in SoNaR provide the different general corpora, each consisting of a subsection of SoNaR (e.g. newspaper articles, Wikipedia, emails, ..). A word is then associated with the corpus where it has the highest relative frequency. For multiple word candidates, the relative frequency is averaged over all words, so if one word in a concept is more related to e.g. Wikipedia articles, but another word is more strongly correlated (i.e. has a larger relative frequency) with the medical corpus, the combination of both words is considered medical. Medical concepts consisting of multiple words are afterwards filtered on their pointwise mutual information score, with a threshold of 0.1. This allows for filtering out accidental, nonsensical concepts (e.g. 'patient cardiogenic shock', which should be two separate concepts). An advantage is that this approach allows for extracting overlapping concepts (e.g. both 'temporary pacemaker' and 'pacemaker' are extracted from the same sentence, improving generalisability for a later feature 'pacemaker'). For LMI-based concepts, we see that on average 39% of the candidate concepts are retained, over all specialties and text types, with an average of 44 concepts per text.

2.3.3. Bootstrapped pattern mining approach (BPM)

Gupta et al [26] introduced a bootstrapped pattern mining method (BPM), as depicted in figure 3. This semi-supervised technique uses a seed list of concepts. This seed list is extracted in four categories (i.e. medication, bodypart, diagnosis/symptom, action) by manually selecting 10 different examples from the unannotated corpus in each category. The seed list is used to learn patterns (e.g. words surrounding those concepts, previously defined as context) iteratively by first finding patterns, looking at the concepts within extracted patterns, and repeating with the new concepts.

As an example, we follow the cycle in figure3. In step 0, the seed term 'diabetes' is considered an example of the diagnosis category, and the seed term 'heart' is defined as a body part. In step 1, patterns are extracted by looking at the words surrounding previously detected concepts (or in the first iteration, the seed terms). In our example, this results in the patterns 'is diagnosed with X' and 'X can

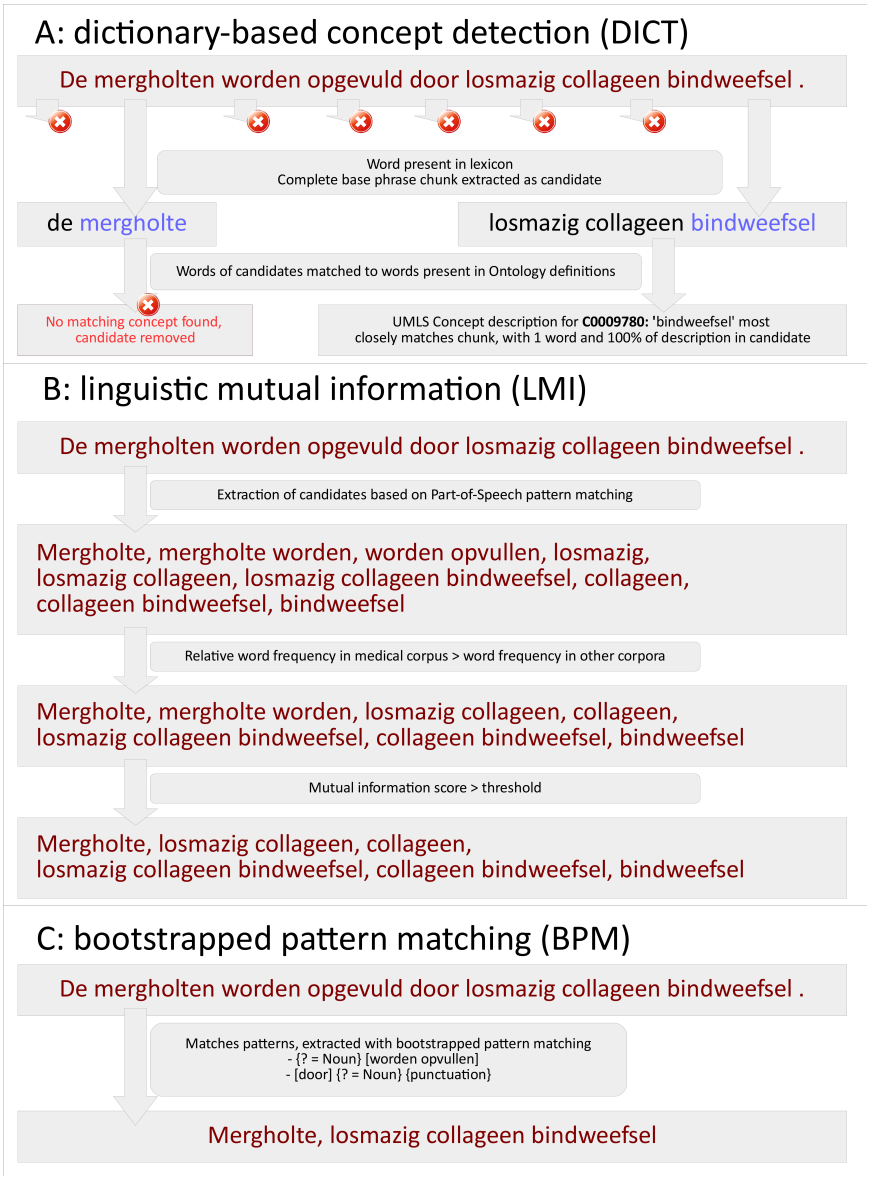

Figure 2: Example of the three concept detection methods. A: dictionary-based approach, B: linguistic PoS and pointwise mutual information method (LMI), C: bootstrapped pattern mining approach. The gloss of the original input sentence is 'The bone marrow cavities are being filled with loose connective collagenous tissue'.

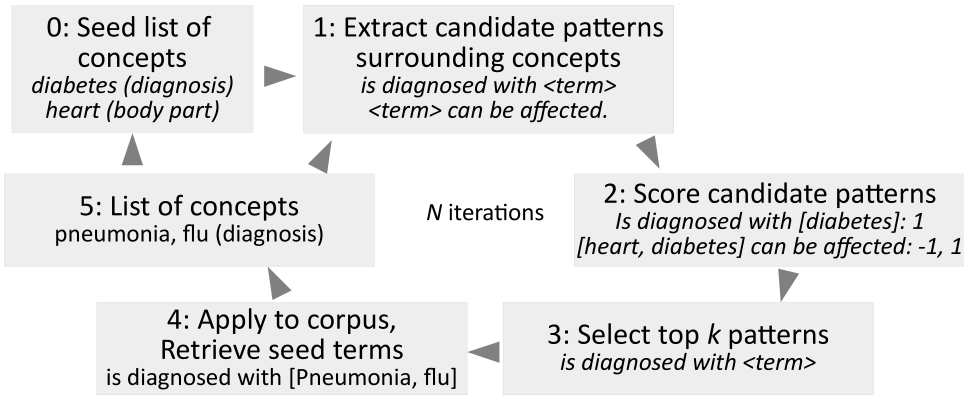

Figure 3: Extraction cycle for patterns surrounding candidate concepts. An example for the first iteration is shown in italics.

be affected' being extracted with the seed term 'diabetes'. In step 2, both patterns are evaluated by considering how often they occur with other seed terms. If they occur more often with seed terms from the correct category, the pattern gains weight, while if they can be associated with seed terms from another category (or negative terms not associated with any category), they lose weight. In the example, the pattern 'X can be affected' could be associated with both the seed term 'diabetes' and 'heart'. In step 3, the best scoring patterns are selected. In our example, we only select one pattern. In step 4, the selected patterns are used to extract new entities for each category, which are appended to the entity list in step 5. The pattern 'is diagnosed with X' could also be applied to the terms 'pneumonia' and 'flu', which are then both added to the seed list.

Both a list of patterns and a list of concepts can be seen as the output of this cycle. In the BPM-based concept detection approach, we opted to apply the patterns to the test data, since this list of patterns results in the same concepts, and some concepts might not have occurred in the unannotated corpus on which the BPM model was trained. We first find patterns on the historical corpus, then apply these patterns in the clinical coding pipeline to extract concepts, as seen in figure 2C.

The complementarity of these methods is evaluated by using two ensemble methods, one applying majority voting (in which a concept needs to be detected by at least two methods), and the other including all concepts from each detection method. In figure 4, the number of detected concepts (on average) per patient stay can be seen.

2.4. Representation of concepts in word2vec

While detected concepts can represent a text accurately, a model does not know which words might be similar to each other and are generalisable towards each other. We try to include this information by constructing distributional semantic representations of concepts (word embeddings) on the basis of an unannotated corpus, using a word2vec model. These representations are used to provide information on the meaning of a concept.

The word2vec model was trained with an historical dataset. Preprocessing included tokenization and lemmatization. Stopwords and function words were not removed from the texts. The individual words of concepts occurring within the text were replaced with concepts detected with the LMI method (see Concept Detection). When concepts overlapped, concepts that had a higher similarity score and concepts that had more words (were more specific) took precedence. We configured the word2vec skip-gram model with a window size of 10 words and an output dimensionality of 300 vectors. These parameters were determined with a grid-search, but the differences between model configurations were insignificant (results not shown).

Four techniques to integrate the word2vec model in the prediction pipeline are presented. The nearest-neighbours method looks for the concepts that are most similar to concepts present in the patient stay. For example, if a concept 'myocarditis' occurs in a patient stay, the n most similar concepts learned by the word2vec model are retrieved. The results (in this case 'carditis', 'inflammatory_cardiomyopathy' and the wrongly spelled 'myocartitis') are then added to the list of features of that patient stay. Cosine distance is used as similarity measure, and the 10 closest neighbours were included as features.

The second method generalizes the concepts by fitting a K-means model over the word2vec model to assign clusters to nodes in the word2vec model, with the advantage of having fewer generated features compared to the neighbour-based approach. For example, if we have two patient stays, where one contains the concept 'myocarditis' and the other contains the concept 'inflammatory_cardiomyopathy', and both concepts belong to the cluster 5012, the feature 'word2vec_cluster_5012' is added to the list of features from each patient stay.

The third technique creates a representation for an entire patient stay by averaging all concept vectors in that patient stay, resulting in a single vector. This vector is then used directly as the feature vector.

In contrast to averaging concept vectors, the fourth technique averages document vectors generated with a paragraph vector model [29]. This model creates representations for a sentence, paragraph, or documents in the same space as the word2vec model by inferring a vector for these documents. This includes an additional training step, in which the words in the document are added to the document vector. While both techniques produce a single vector, averaged concept vectors are only influenced by the detected concepts, while averaged document vectors embed a representation of entire texts.

2.5. Representation of concepts using UMLS metathesaurus

The UMLS metathesaurus is also a great tool to generalise concepts and find similar concepts [8]. It can be used to expand a list of terms with related terms, when a concept is already linked to a thesaurus. The first UMLS-based method retrieves synonyms and parent concepts for dictionary-based concepts (as these are already linked to UMLS), and adds these extra concepts as neighbouring nodes.

For LMI-based concepts, a direct relation to UMLS would restrict these concepts to (a subset) of the dictionary-based concepts, which limits the usability of UMLS for synonym expansion. By first expanding the LMI-based concepts with neighbouring concepts retrieved through word2vec, a much larger list of concepts arises.

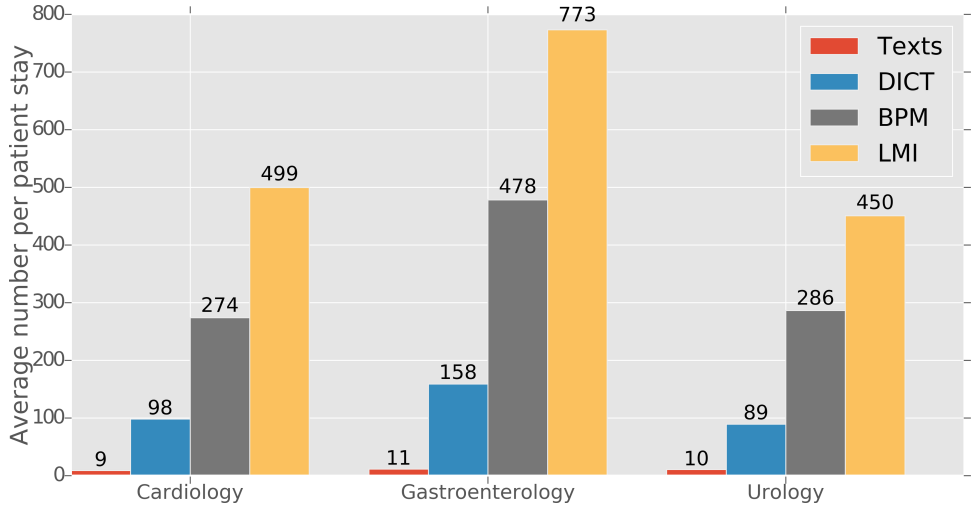

Figure 4: This figure shows the average number of texts and the number of extracted concepts per patient stay (on the y-axis) in each disease area. Texts indicates the number of texts, DICT indicates the number of dictionary-based concepts, BPM indicates the number of bootstrapped-pattern matching concepts and LMI indicates the number of linguistic-PoS and pointwise mutual information-based concepts.

This list of concepts can then be restricted to concepts that can be related to UMLS. On average, this causes only 21% of the word2vec neighbours to be retained, using the same settings as the word2vec neighbours model to retrieve them.

2.6. Predicting clinical codes

Each model is evaluated on its results when predicting clinical codes for a set of patient stays. We treated the task as a multi-label classification task (with an individual classifier for each clinical code) in which a single patient stay might have zero, one or more assigned codes. Ten-fold cross-validation was applied. We used Random Forests (with a maximum of 100 trees) as a classifier [30]. Feature selection for the BoW representation consists of TF-IDF [10], no feature selection was used for the averaged word2vec vector, and for all other approaches, mRMR was used [31]. Features are selected by the mRMR method based on a maximal distance between features (minimizing redundancy) and still retaining a high correlation to the classes (maximizing relevance).

For each patient stay, a list of concepts is derived from all associated texts (which can be zero, one or more) with the methods as presented in the Concept Detection section. This list of concepts is then used or supplemented with features derived from the word2vec model to predict clinical codes.

2.7. Evaluation

The experiments are evaluated with micro-averaged F-measure, exact match, and macro-averaged F-measure. Micro-averaged F-measure sums up individual predictions, without weighing them based on samples or labels. F-measure is the harmonic mean of precision and recall and is an indicator for the overall predictive power of models. Sample-based evaluation is performed using Exact match and F-measure, macro-averaged over each sample. Each sample, regardless of the number of codes assigned to it, contributes equally to the total score. Exact match is a strict version of accuracy that only considers a sample where all classes are correctly predicted as a true positive.

The used models return class probabilities, which can be used for further tuning to increase recall or precision, depending on the priorities of the task at hand. This further tuning is not carried out in this research, as the models are already optimized for F-measure within the cross validation loop.

3. Results

The experiments are presented separately for different disease areas as well as for procedural and diagnostic ICD-9-CM codes. The significance of the results is demonstrated by using the technique of approximate random testing [32]. This method measures the similarity of the results on a per-sample basis between two systems. A similarity under 5% indicates that it is 95% likely that two systems are independent. Complete significance score sheets against the baseline and a partial sheet against LMI concepts can be found in supplementary.

3.1. Bag-of-words (BoW) baseline

The baseline for clinical code assignments is calculated by using a BoW over all texts associated with a patient stay. The words in the BoW are filtered on the TF-IDF score of the individual words. 3,000 words with the highest TF-IDF score are directly used as features. This threshold was determined as the overall optimal parameter with a grid search.

	Card,Diag	Card,Proc	Gast, Diag	Gast,Proc	Uro,Diag	Uro,Proc
Baseline (bow of all texts)	36.6	69.0	22.0	49.1	25.3	46.5
Concept detection methods						
Dictionary-based concepts	38.5	77.7	**23.1**	**65.3**	23.5	55.9
BPM-based concepts	37.3	82.0	19.9	58.8	21.4	55.3
LMI-based concepts	42.0	83.2	22.1	58.6	25.8	**58.3**
Majority-voted concepts	39.4	79.3	22.6	64.2	21.5	51.2
Bag of concepts	**42.6**	**84.8**	21.8	57.6	**26.7**	56.7
Methods for representation of concepts in word2vec						
Word2Vec Neighbours with LMI-based concepts	**43.8**	82.8	**25.0**	55.2	**30.6**	52.3
Word2Vec Cluster with LMI-based concepts	42.2	**82.9**	22.3	**57.5**	26.1	**53.2**
Average Word2Vec-vector with LMI-based concepts	36.1	72.7	19.9	54.9	25.7	53.1
Average Document vector	31.4	52	10.8	23.2	15.6	26.3
Methods for representation of concepts in UMLS						
UMLS neighbours with dictionary based concepts	43.4	78.8	**30.7**	65.7	27.5	48.4
Word2Vec neighbours, restricted with UMLS, with LMI-based concepts	**44.8**	**82.7**	28.9	**67.1**	**30.5**	**60.9**

Table 2: micro-averaged F-measure (in%) for each disease area and code type for the different concept representation methods. The best result in each category is marked in bold. The first element in the column name stands for the disease area (card, gast and uro are short for cardiology, gastroenterology and urology) while the second element stands for the predicted codeset (diag are diagnostic ICD-9-CM codes and proc are procedural ICD-9-CM codes).

3.2. Raw concepts as features

In table 2, the rows under the header 'Concept detection method' represent concepts directly used as features. Each row represents a different method by which concepts have been extracted as features in the machine learning setup. Complete overviews can be found in the supplementary materials. The underlying recall and precision of the presented F-measure scores were balanced, with no real outliers in either of them. In general, we see that using a list of concepts systematically outperforms using the bag-of-words baseline approach when predicting clinical codes, with the exception of diagnostic codes for urology. The dictionary-based approach (DICT) performs well and is the best source for prediction in the disease area of gastroenterology. With the linguistic pattern matching (LMI) approach, the results obtained outperformed the dictionary-based approach for the fields of cardiology and urology. The systems used for the dictionary-based approach were always significantly different from systems used for the LMI-based approach, ranging from 0.2% similarity when predicting procedure codes for urology to 3.2% similarity when predicting diagnostic codes for cardiology [32]. The first ensemble method, a bag of all concepts, is on par with the best individual concept detection method, with the best result for two of the experiments. Majority voting performs worse than individual predictions for each experiment. Bootstrapped pattern mining (BPM) had an F-measure that is in general lowest of the concept detection methods. The poorer results for the LMI approach in the field of gastroenterology can be explained by an overflow of features, introducing more noise. Figure 4 shows that LMI detects more concepts for gastroenterology (773) than for urology (450) or cardiology (499). This is further confirmed by poorer results for the bag of concepts method, and better results for the majority vote concepts, which filter out a large number of the extra (obsolete) features introduced by LMI.

3.3. Adding word2vec-based features to raw concepts

In table 2, rows indicated with 'Word2Vec clusters with x', the input space consists of the concepts and cluster ids from a K-means clustering model on top of the word2vec model. 'Word2Vec neighbours with x' have the concepts and concept nodes determined as neighbours in the word2vec model as features. For readability, we only included the best concept detection methods in this table, the rest can be found in supplementary materials. When predicting diagnostic codes with word2vec-neighbours in addition to LMI-based concepts, the F-measure increases with 1 to 4%. The similarity between predictions with word2vec-neighbours and LMI-concepts ranges from 0.3% similarity for diagnostic codes in gastrology to 1.6% similarity for procedural codes in cardiology [32]. When predicting procedural codes, the F-measure decreased. Including word2vec-cluster ids as features do not increase the f-measure.

3.4. Patient stay representation through a single vector

In table 2, the row indicated with 'Averaged word2vec vector with LMI-concepts' represent the results achieved when an averaged-out word2vec vector (averaged for all concepts) is used to determine clinical codes. The row indicated with 'Averaged document vector' represents the results when using document vectors (averaged for each text related to a patient stay). Both setups perform worse when evaluating with micro-averaged F-measure, with an average vector on the concept-level scoring better than an average vector on the document-level.

3.5. Using UMLS to enhance representations

The row 'UMLS neighbours with dictionary-based concepts' in table 2 show the results when dictionary-based concepts were expanded with parent and synonym concepts retrieved directly in UMLS. The micro-averaged F-measure is consistently higher than solely using dictionary-based concepts without expansion, except for procedure

codes in urology. For diagnostic codes in gastrology, this method yields the best results of al variants.

The row indicated with 'Word2Vec neighbours, restricted with UMLS, with LMI-based concepts' corresponds to the experiments where LMI-concepts, expanded with neighbouring concepts determined by word2vec, which were then restricted to only contain concepts that could be linked back to UMLS. The micro-averaged F-measure for these experiments is best, except for procedure codes in cardiology.

3.6. Sample-based evaluation

Sample-based evaluation looks at the results differently, to give a viewpoint based on how well the algorithms perform for each sample (i.e. patient stay) instead of for each predicted code as is the case in a micro-averaged evaluation. This lowers the total weight of frequently occurring codes. Exact match reflects the total number of samples that were correctly predicted, macro-averaged F-measure (over the sample) weighs predictions so that each sample has an equal contribution.

In Table 3, we see that only a fraction of the samples for diagnostic codes are predicted correctly, showing no real difference between any of the techniques. The high(er) baseline here for diagnostic codes in gastroenterology is explained by exact match considering stays where no codes were assigned (and predicted) as a true positive. For procedure codes, we see an improvement in exact match with the same trends as observed for micro-averaged F-measure: in gastroenterology, dictionary-based concepts score best, while for cardiology and urology, LMI-based concepts and a bag of all concepts score better.

Macro-averaged F-measure, as seen in Table 4, shows finer distinctions, as predictions are weighted by the total number of gold-standard codes in a sample (and still influence the results when only part of the sample is predicted correctly). A correct prediction in a sample with 5 gold-standard codes thus has a lower weight than a correct prediction in a sample with 3 gold-standard codes. Trends in the results for concept detection methods stay approximately the same when comparing micro- and macro-averaged F-measure, although procedure codes in gastroenterology increased for LMI-based concepts. We also see that averaging the word2vec vectors from concepts performs better on macro-averaged F-measure. When predicting diagnostic codes in urology, the results for all methods, except majority-voted concepts, drops below baseline. The macro-averaged F-measure for the UMLS-restricted word2vec neighbours model never reach top performance.

4. Discussion

The approach of medical information extraction (i.e. concept detection) presented here differs from previous work, as we do not evaluate the extracted information directly, but rather look at the results achieved by using this information to solve a specific task (i.e. the prediction of clinical codes). While we cannot be certain that the extracted concepts cover the entire set of concepts available in the texts, we show that this pragmatic detection of concepts has a positive influence on the task. Furthermore, while Named Entity Recognition (NER)/concept detection is often treated as a task on its own, in practice, it will always be used as a means to perform a certain task. The results we present show the usability of our approach in those cases.

The results show that representing clinical texts as a list of concepts reduces noise and enables the retrieval of extra information for our specific task. While both LMI and bag of concepts carry concepts that are not very informative, the feature selection and classification methods used handle this well (except for gastroenterology). With LMI, we see that selected concepts often also contain commonly misspelled terms, abbreviations and definitions that are not present in dictionaries. The results in gastroenterology show that LMI is not always able to outperform a dictionary-based approach. We expect this is due to an overload of features being added, which is further confirmed by better performance when using majority-based voting and worse performance when using a bag of concepts (where concepts from all different techniques are directly added as features).

One of the reasons for the overload of features is that we opted to restrict the boundaries for candidate concepts using part-of-speech tags rather than using dependency parsing. While this would allow us to further restrict the number of candidates, it also requires an set of NLP modules specifically trained on clinical data. In our case, this restriction would cause a large percentage of relevant candidates to be removed. This choice resulted in an increase in the number of both correct and erroneous concepts.

Experiments with features found by distributional semantic models produce mixed results. An improvement is mainly seen when we are predicting diagnostic codes. We hypothesize that in texts, the mention of diagnoses is more vague than mentions of procedures. While a procedure was either performed or not, a diagnosis often has a tentative nature. The terminology used for describing a diagnosis or a disease is also wider. By adding concepts that are considered similar, the model succeeds indeed in finding similarity between documents describing patient stays where there were no matching features before. Additionally, because ICD-9-CM diagnostic codes are more fine-grained than ICD-9-CM procedure codes, the features are also required to reflect those details. The main disadvantage of using a distributional semantic model to generate features is the abundance of generated features.

Restricting the neighbours retrieved with word2vec by selecting only the ones that can be related to an UMLS-concept integrates expert knowledge into the model. This

Table 3: Exact match (in %) for each disease area and code type for the different concept representation methods. The first element in the column name stands for the disease area (card, gast and uro are short for cardiology, gastroenterology and urology) while the second element stands for the predicted codeset (diag are diagnostic ICD-9-CM codes and proc are procedural ICD-9-CM codes).

	Card, Diag	Card, Proc	Gast, Diag	Gast, Proc	Uro, Diag	Uro, Proc
Baseline (bow of all texts)	5.7	33.3	33.9	54.8	8.4	37.6
Concept detection methods						
Dictionary-based concepts	5.7	42.8	33.9	**66.7**	9	44.2
BPM-based concepts	5.6	52.5	33.9	63.2	8.6	45.2
LMI-based concepts	6.1	57.2	33.9	62.4	**9.5**	**46.3**
Majority-voted concepts	5.9	47.8	33.9	66.3	9.1	38.4
Bag of concepts	**6.2**	**58.9**	33.9	60.7	9.3	46.2
Methods for representation of concepts in word2vec						
Word2Vec neighbours with LMI-based concepts	**6.6**	55.6	33.9	60	**10**	38.7
Word2Vec cluster with LMI-based concepts	6.3	**56**	33.9	**61.3**	9.4	41.1
Average Word2Vec-vector with LMI-based concepts	5.7	39.8	33.9	58.7	8.6	**42.9**
Average Document vector	5.6	24	33.9	42.1	1.3	24.9
Methods for representation of concepts in UMLS						
UMLS neighbours with dictionary based concepts	6.2	45.1	**34.1**	67.6	9.6	36.4
Word2Vec neighbours, restricted with UMLS, with LMI-based concepts	**6.8**	**56.6**	**34.1**	**67.9**	**10.1**	**48.9**

Table 4: Macro-averaged F-measure (in %) over samples for each disease area and code type for the different concept representation methods. The first element in the column name stands for the disease area (card, gast and uro are short for cardiology, gastroenterology and urology) while the second element stands for the predicted codeset (diag are diagnostic ICD-9-CM codes and proc are procedural ICD-9-CM codes).

	Card, Diag	Card, Proc	Gast, Diag	Gast, Proc	Uro, Diag	Uro, Proc
Baseline (bow of all texts)	37.6	70	15.3	44	36.3	30
Concept detection methods						
Dictionary-based concepts	40.3	70.7	**19.8**	**50**	31.7	**35**
BPM-based concepts	38.5	73.4	16	45	25.7	**35**
LMI-based concepts	41.6	73.1	18.2	**50**	28.5	31.7
Majority-voted concepts	38.8	69.7	18.3	45	**39.8**	33.3
Bag of concepts	**44.9**	**75.3**	18	46.7	28.8	33.3
Methods for representation of concepts in word2vec						
Word2Vec neighbours with LMI-based concepts	**46.9**	**73.3**	20.7	40	32.5	28.3
Word2Vec cluster with LMI-based concepts	40.7	71.2	20	**50**	26.8	33.3
Word2Vec average vector with LMI-based concepts	36	69.1	**21**	43.3	**33.5**	**36.7**
Average Document vector	35	62	10.2	19.5	27.7	29.9
Methods for representation of concepts in UMLS						
UMLS neighbours with dictionary based concepts	**44.1**	70.4	20.2	**50**	**36.1**	31.7
Word2Vec neighbours, restricted with UMLS, with LMI-based concepts	41	**73.6**	**23.0**	46.7	35.5	**32**

greatly reduced the number of features, which consistently improved the results over both using LMI-based concepts as features and using LMI-based concepts and word2vec neighbours in combination as features. The performance of dictionary-based concepts in gastroenterology is also outperformed in this model, which can be an effect of both the expert knowledge being integrated and the reduced number of features. By using UMLS to restrict related concepts, we avoid directly assigning expert-based information (and missing concepts that are not present in this ontology), and are still able to have relations to these concepts that are based on expert knowledge.
Averaging the concept vectors generated with word2vec for all concepts in a stay seems advantageous when evaluating on a per-sample basis, but not when evaluating per individual code prediction. This might indicate that the vectors represent the concepts well, but generalize them more strongly and lose nuances that might be important to predict frequent codes. Averaging vectors inferred on the document level do not provide a useful representation for this task. This indicates that the documents cannot be adequately captured using a single vector, mainly because all information in a document needs to be reflected in this vector (not only medical concepts). By averaging the vectors for multiple documents, this effect is amplified.

A downside to LMI concepts (or any concepts not defined in a dictionary) is that there is no expert-annotated context. For example, if 'carditis' is detected through its direct definition in UMLS (C0869523), there is an expert interpretation linked to the concept, which makes it easier to find related concepts. While using a distributional semantics model also adds an interpretation and metadata to the concepts, the results show that this is still an error-prone process. A manual check of the returned neighbours for certain concepts yielded some surprisingly accurate results, but also some disappointing ones. This is consistent with a distributional semantics model relying on the context in which words occur, and the context of two words can be similar even if the concepts in question are not.

Due to the use of a real-life dataset, we were dependent on the code system present in this data (i.e. ICD-9-CM). While this has an effect on the portability to other systems, the proposed methodology is directly applicable to other coding systems. The relatively low F-measures seen in the results are an indication of the complexity of the task. This is reflected in other studies, where codes predicted on real-life datasets tend to achieve low F-measures [22, 24]. A benefit of the dataset used is that the results are more indicative of how the algorithms behave in a practical context.

When clinical code prediction is used in an application, optimizing the output of the model for either precision (which allows a classifier to automatically assign clinical codes on a small portion of the files) or for recall (which can be used in a setup to assist clinical coders in retrieving the correct codes for a patient file) is a necessary step. This choice was not made in our evaluation, because we were not directly evaluating the efficacy of the algorithm in a clinical coding environment and wanted to show an overall picture of how the model performs.

4.1. Future work

To further explore the value of document representations, we could use documents directly to represent a stay, instead of using concepts. Although documents are unique to one patient stay, adding the nearest neighbouring documents would create the overlap required to assign classes to stays. If documents are more similar to each other, we would then expect the classes to be more similar as well. This technique requires a higher quality of document vectors than we have been able to create.

5. Conclusion

In this study, we introduced methods for medical concept detection and concept representation. We compared the results of concepts learned with an unsupervised technique (LMI) with using concepts derived from expert-based dictionaries (DICT) in the task of clinical code assignment. Both diagnostic and procedural codes were predicted, for three different medical fields with a varying dataset size. Both the DICT and LMI methods achieved a higher F-measure than a bag-of-words approach, with LMI-based concepts performing best in general. These results confirm that the concepts succeed in capturing more information, while reducing the noise present. LMI turns out to be a viable alternative method to retrieve concepts in text when expert-developed dictionaries are not (or only partly) available (as is the case for Dutch, for instance), given that a sufficient amount of raw/unannotated data is available.

In addition, a distributional semantics model (DSM) was used to interpret the meaning of the concepts based on the context in which they occur. Three different methods of introducing the extra knowledge were compared: introducing similar concepts in a patient stay with the nearest-neighbours method, introducing cluster ids of similar concepts with the cluster-method, and directly using the vectors learned by the DSM to position a patient stay. The nearest-neighbours method showed an improved F-measure when predicting diagnostic codes, for each medical field. This improvement was not seen when predicting procedure codes. By restricting the nearest-neighbours method by only retaining neighbouring concepts that can be related to the UMLS, the results are improved further, for both diagnostic and procedure codes.

6. Acknowledgements

This work was supported by the Agency for Innovation by Science and Technology in Flanders (IWT) grant number 131137.

References

[1] P. R. Payne, Biomedical knowledge integration, PLoS Comput Biol 8 (12).

[2] C.-J. Hsiao, E. Hing, et al., Use and Characteristics of Electronic Health Record Systems Among Office-Based Physician Practices, United States, 2001-2012, US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.

[3] J. J. Cimino, Improving the electronic health recordare clinicians getting what they wished for?, JAMA 309 (10) (2013) 991–992.

[4] G. S. Alotaibi, C. Wu, A. Senthilselvan, M. S. McMurtry, The validity of icd codes coupled with imaging procedure codes for identifying acute venous thromboembolism using administrative data, Vascular Medicine.

[5] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, W. F. Stewart, Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, International journal of medical informatics 83 (12) (2014) 983–992.

[6] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, J. C. Denny, Medex: a medication information extraction system for clinical narratives, Journal of the American Medical Informatics Association 17 (1) (2010) 19–24.

[7] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman, G. Savova, Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, Journal of the American Medical Informatics Association 22 (1) (2015) 143–154.

[8] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (suppl 1) (2004) D267–D270.

[9] B. Tang, Y. Wu, M. Jiang, J. C. Denny, H. Xu, Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model., in: CLEF (Working Notes), 2013.

[10] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (5) (1988) 513–523.

[11] J. Pathak, K. R. Bailey, C. E. Beebe, S. Bethard, D. S. Carrell, P. J. Chen, D. Dligach, C. M. Endle, L. A. Hart, P. J. Haug, et al., Normalization and standardization of electronic health records for high-throughput phenotyping: the sharpn consortium, Journal of the American Medical Informatics Association 20 (e2) (2013) e341–e348.

[12] P. D. Turney, P. Pantel, et al., From frequency to meaning: Vector space models of semantics, Journal of artificial intelligence research 37 (1) (2010) 141–188.

[13] S. Jonnalagadda, T. Cohen, S. Wu, G. Gonzalez, Enhancing clinical concept extraction with distributional semantics, Journal of biomedical informatics 45 (1) (2012) 129–140.

[14] P. Kanerva, J. Kristofersson, A. Holst, Random indexing of text samples for latent semantic analysis, in: Proceedings of the 22nd annual conference of the cognitive science society, Vol. 1036, Citeseer, 2000.

[15] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, M. Duneld, Synonym extraction and abbreviation expansion with ensembles of semantic spaces., J. Biomedical Semantics 5 (6).

[16] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, In Proceedings of Workshop at the International Conference on Learning Representations.

[17] H. Moen, F. Ginter, E. Marsi, L.-M. Peltonen, T. Salakoski, S. Salanterä, Care episode retrieval: distributional semantic models for information retrieval in the clinical domain, BMC medical informatics and decision making 15 (Suppl 2) (2015) S2.

[18] W. H. Organization, et al., International classification of diseases (icd).

[19] T. Mikolov, word2vec: Tool for computing continuous distributed representations of words (2013).

[20] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, W. R. Hersh, A systematic literature review of automated clinical coding and classification systems, Journal of the American Medical Informatics Association 17 (6) (2010) 646–651.

[21] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, W. Duch, A shared task involving multi-label classification of clinical free text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, 2007, pp. 97–104.

[22] A. Perotte, R. Pivovarov, K. Natarajan, N. Weiskopf, F. Wood, N. Elhadad, Diagnosis code assignment: models and evaluation metrics, Journal of the American Medical Informatics Association 21 (2) (2014) 231–237.

[23] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, R. G. Mark, Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database, Critical care medicine 39 (5) (2011) 952.

[24] E. Scheurwegs, K. Luyckx, L. Luyten, W. Daelemans, T. Van den Bulcke, Data integration of structured and unstructured sources for assigning clinical codes to patient stays, Journal of the American Medical Informatics Association`doi: 10.1093/jamia/ocv115`.

[25] A. v. d. Bosch, B. Busser, S. Canisius, W. Daelemans, An efficient memory-based morphosyntactic tagger and parser for dutch, LOT Occasional Series 7 (2007) 191–206.

[26] S. Gupta, C. D. Manning, Improved pattern learning for bootstrapped entity extraction, in: CoNLL, 2014, pp. 98–108.

[27] F. A. voor Geneesmiddelen en Gezondheidsproducten (FAGG), Belgian and european lists of medication brands and active compounds, `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm`, accessed: 2016-02-19.

[28] N. Oostdijk, M. Reynaert, P. Monachesi, G. Van Noord, R. Ordelman, I. Schuurman, V. Vandeghinste, From d-coi to sonar: a reference corpus for dutch., in: LREC, 2008.

[29] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, arXiv preprint arXiv:1405.4053.

[30] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[31] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (8) (2005) 1226–1238.

[32] A. Yeh, More accurate tests for the statistical significance of result differences, in: Proceedings of the 18th conference on Computational linguistics-Volume 2, Association for Computational Linguistics, 2000, pp. 947–953.