Care Episode Retrieval

Hans Moen[1], Erwin Marsi[1], Filip Ginter[2],
Laura-Maria Murtola[3,4], Tapio Salakoski[2], Sanna Salanterä[3,4],
[1]Dept. of Computer and Information Science,
Norwegian University of Science and Technology, Norway
[2]Dept. of Information Technology, University of Turku, Finland
[3]Dept. of Nursing Science, University of Turku, Finland
[4]Turku University Hospital, Finland
{hans.moen,emarsi}@idi.ntnu.no, ginter@cs.utu.fi,
{lmemur,tapio.salakoski,sansala}@utu.fi

Abstract

The documentation of a care episode consists of clinical notes concerning patient care, concluded with a discharge summary. Care episodes are stored electronically and used throughout the health care sector by patients, administrators and professionals from different areas, primarily for clinical purposes, but also for secondary purposes such as decision support and research. A common use case is, given a – possibly unfinished – care episode, to retrieve the most similar care episodes among the records. This paper presents several methods for information retrieval, focusing on care episode retrieval, based on textual similarity, where similarity is measured through domain-specific modelling of the distributional semantics of words. Models include variants of *random indexing* and a semantic neural network model called *word2vec*. A novel method is introduced that utilizes the ICD-10 codes attached to care episodes to better induce domain-specificity in the semantic model. We report on an experimental evaluation of care episode retrieval that circumvents the lack of human judgements regarding episode relevance by exploiting (1) ICD-10 codes of care episodes and (2) semantic similarity between their discharge summaries. Results suggest that several of the methods proposed outperform a state-of-the art search engine (Lucene) on the retrieval task.

1 Introduction

Information retrieval (IR) aims at retrieving and ranking documents relative to a textual query expressing the information need of a user (Manning et al., 2008). IR has become a crucial technology for many organisations that deal with vast amounts of partly structured and unstructured (free text) data stored in electronic format, including hospitals and other health care providers. IR is an essential part of the clinical practice; e.g., on-line IR systems are associated with substantial improvements in clinicians decision-making concerning clinical problems (Westbrook et al., 2005).

The different stages of the clinical care of a patient are documented in *clinical care notes*, consisting mainly of free text. A *care episode* consists of a sequence of individual clinical care notes, concluded by a discharge summary, as illustrated in Figure 1. Care episodes are stored in electronic format in *electronic health record* (EHR) systems. These systems are used throughout the health care sector by patients, administrators and professionals from different areas, primarily for clinical purposes, but also for secondary purposes such as decision support and research (Häyrinen et al., 2008). IR from EHR in general is therefore a common and important task.

This paper focuses on the particular task of retrieving those care episodes that are most similar to the sequence of clinical notes for a given patient, which we will call *care episode retrieval*. In conventional IR, the query typically consists of several keywords or a short phrase, while the retrievable units are typically documents. In contrast, in care episode retrieval, the query consist of the clinical notes contained in a care episode. The discharge summary is used separately for evalu-

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 116–124,
Gothenburg, Sweden, April 26-30 2014.

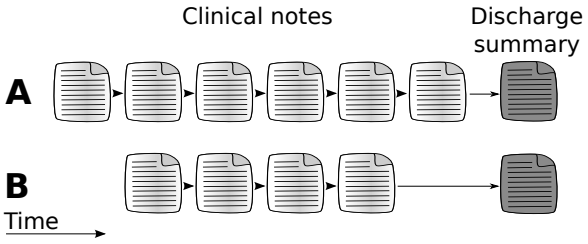

Figure 1: Illustration of care episode retrieval. The two care episodes (A and B) are composed of a number of individual clinical notes and a single discharge summary. Given an ongoing care episode (minus the discharge summary), the task is to retrieve other, similar care episodes.

ation purposes, and is assumed to be unavailable for constructing a query at retrieval time. Retrievable units are thus complete care episodes without summaries.

We envision a number of different use cases for a care episode retrieval system. Firstly, it could facilitate clinicians in decision-making. For example, given a patient that is being treated in a hospital, an involved clinician may want to find previous patients that are similar in terms of their health history, symptoms or received treatments. Supplementary input from the clinician would enable the system to give heightened weight to keywords of particular interest within the care episodes, which would further be emphasized in the semantic similarity calculation during IR. It may help considerably to see what similar patients have received in terms of medication and further treatment, what related issues such as bi-conditions or risks occurred, how other clinicians have described certain aspects, what clinical practice guidelines have been utilized, and so on. This relates to the underlying principle in textual case-based reasoning (Lenz et al., 1998). Secondly, it could help management to get almost real time information concerning the overall situation on the unit for a specific follow-up period. Such a system could for example support managerial decision-making with statistical information concerning care trends on the unit, adverse events or infections. Thirdly, it could facilitate knowledge discovery and research. For instance, it could enable researchers to map or cluster similar care episodes to find common symptoms or conditions. In sum, care episode retrieval is likely to improve care quality and consistency in hospitals.

From the perspective of NLP, care episode retrieval – and IR from EHRs in general – is a challenging task. It differs from general-purpose web search in that the vocabulary, the information needs and the queries of clinicians are highly specialised (Yang et al., 2011). Clinical notes contain highly domain-specific terminology (Rector, 1999; Friedman et al., 2002; Allvin et al., 2010) and generic text processing resources are therefore often suboptimal or inadequate (Shatkay, 2005). At the same time, development of dedicated clinical NLP tools and resources is often difficult and costly. For example, popular data-driven approaches to NLP are based on supervised learning, which requires substantial amounts of tailored training data, typically built through manual annotation by annotators who need both linguistic and clinical knowledge. Additionally, variations in the language and terminology used in sub-domains within and across health care organisations greatly limit the scope of applicability of such training data (Rector, 1999).

Recent work has shown that distributional models of semantics, induced in an unsupervised manner from large corpora of clinical and/or medical text, are well suited as a resource-light approach to capturing and representing domain-specific terminology (Pedersen et al., 2007; Koopman et al., 2012; Henriksson et al., 2014). This raises the question to what extent distributional models of semantics can alleviate the aforementioned problems of NLP in the clinical domain. The work reported here investigates to what extent distributional models of semantics, built from a corpus of clinical text in an fully unsupervised manner, can be used for care episode retrieval. Models include several variants of *random indexing* and a semantic neural network model called *word2vec*, which will be described in more detail in Section 4.

It has been argued that clinical NLP should exploit existing knowledge resources such as knowledge bases about medications, treatments, diseases, symptoms and care plans, despite these not having been explicitly built for doing clinical NLP (Friedman et al., 2013). Along these lines, a novel method is proposed here that utilizes the ICD-10 codes – diagnostic labels attached to care episodes by clinicians – to better induce domain-specificity in the semantic model. Experimental results suggest that this method outperforms a state-of-the art search engine (Lucene) on the task of care episode

retrieval.

Apart from issues related to clinical terminology, another problem in care episode retrieval is the lack of benchmark data, such as the relevance scores produced by human judges commonly used for evaluation of IR systems. Although collections of care episodes may be available, producing gold standard similarity scores required for evaluation is costly. Another contribution of this paper is the proposal of evaluation procedures that circumvent the lack of human judgements regarding episode similarity. This is accomplished by exploiting either (1) ICD-10 codes of care episodes or (2) semantic similarity between their discharge summaries. Despite our focus on the specific task of care episode retrieval, we hypothesize that the methods and models proposed here have the potential to increase performance of IR on clinical text in general.

2 Data

The data set used in this study consists of the electronic health records from patients with any type of heart related problem that were admitted to one particular university hospital in Finland between the years 2005-2009. Of these, only the clinical notes written by physician are used. A supporting statement for the research was obtained from the Ethics Committee of the Hospital District (17.2.2009 §67) and permission to conduct the research was obtained from the Medical Director of the Hospital District (2/2009). The total set consist of 66884 care episodes, which amounts to 398040 notes and 64 million words in total. This full set was used for training of the semantic models. To make the experimentation more convenient, we chose to use a subset for evaluation. This comprises 26530 care episodes, amounting to 155562 notes and 25.7 million words in total.

Notes are mostly unstructured, consisting of free text in Finnish. Some meta-data – such as names of the authors, dates, wards, and so on – is present, but is not used for retrieval.

Care episodes have been manually labeled according to the 10th revision of the International Classification of Diseases (ICD-10) (World Health Organization and others, 2013), a standardised tool of diagnostic codes for classifying diseases. Codes are normally applied at the end of the patient's stay, or even after the patient has been discharged from the hospital. Care episodes have one primary ICD-10 code attached and optionally a number of additionally relevant codes. In this study, only the primary one is used, because extraction of the secondary codes is non-trivial.

ICD-10 codes have an internal structure that reflects the classification system ranging from broad categories down to fine-grained subjects. For example, the first character (J) of the code J21.1 signals that it belongs to the broad category *Diseases of the respiratory system*. The next two digits (21) classify the subject as belonging to the subcategory *Acute bronchiolitis*. Finally, the last digit after the dot (1) means that it belongs to the sub-subclass *Acute bronchiolitis due to human metapneumovirus*. There are 356 unique "primary" ICD-10 codes in the evaluation data set.

3 Task

The task addressed in this study is retrieval of care episodes that are similar to each other. In contrast to the normal IR setting, where the search query is derived from a text stating the user's information need, here the query is based on another care episode, which we refer to as the *query episode*. As the query episode may document ongoing treatment, and thus lack a discharge summary and ICD-10 code, neither of these information sources can be relied upon for constructing the query. The task is therefore to retrieve the most similar care episodes using only the information contained in the free text of the clinical notes in the query episode.

Evaluation of retrieval results generally requires an assessment of their relevancy to the query. Since similarity judgements by humans are currently lacking, and obtaining these is time-consuming and costly, we explored alternative ways of evaluating performance on the task. The first alternative is to assume that care episodes are similar if they have the same ICD-10 code. That is, a retrieved care episode is considered correct if its ICD-10 code is identical to the code of the query episode. It should be noted that ICD-10 codes are not used in the query in any of the experiments.

Closer inspection shows that the free text content in care episodes with the same ICD-10 code is indeed quite similar in many cases, but not always. Considering all of them equally similar amounts to an arguably coarse approximation of relevance. The second alternative tries to remedy this issue by measuring the similarity between dis-

charge summaries. That is, if the discharge summary of a retrieved episode is semantically similar to the discharge summary of the query episode, the retrieved episode is assumed to be correct. In practice, textual similarity between discharge summaries, and therefore the relevance score, is continuous rather than binary. It is measured using the same models of distributional semantics used for retrieval, which will be described in Section 4. It should be stressed that the discharge summaries are not taken into consideration during retrieval in any of the experiments and are only used for evaluation.

4 Method

4.1 Semantic models

A crucial part in retrieving similar care episodes is having a good similarity measure. Here similarity between care episodes is measured as the similarity between the words they contain (see Section 4.2). Semantic similarity between words is in turn measured through the use of word space models (WSM), without performing an explicit query expansion step. Several variants of these models were tested, utilizing different techniques and parameters for building them. The models trained and tested in this paper are: (1) classic random indexing with a sliding window using term index vectors and term context vectors (RI-Word); (2) random indexing with index vectors for documents (RI-Doc); (3) random indexing with index vectors for ICD-10 codes (RI-ICD); (4) a version of random indexing where only the term index vectors are used (RI-Index); and (5) a semantic neural network model, using *word2vec* to build word context vectors (Word2vec).

RI-Word

Random Indexing (RI) (Kanerva et al., 2000) is a method for building a (pre) compressed WSM with a fixed dimensionality, done in an incremental fashion. RI consist of the following two steps: First, instead of allocating one dimension in the multidimensional vector space to a single word, each word is assigned an "index vector" as its unique signature in the vector space. Index vectors are generated vectors consisting of mostly zeros together with a randomly distributed set of several 1's and -1's, uniquely distributed for each unique word; The second step is to induce "context vectors" for each word. A context vector represents the *contextual meaning* of a word in the WSM. This is done using a sliding window of a fixed size to traverse a training corpus, inducing context vectors for the center/target word of the sliding window by summing the index vectors of the neighbouring words in the window.

As the dimensionality of the index vectors is fixed, the dimensionality of the vector space will not grow beyond the size $W \times Dim$, where W is the number of unique words in the vocabulary, and Dim being the pre-selected dimensionality to use for the index vectors. As a result, RI models are significantly smaller than plain word space models, making them a lot less computationally expensive. Additionally, the method is fully incremental (additional training data can be added at any given time without having to retrain the existing model), easy to parallelize, and scalable, meaning that it is fast and can be trained on large amounts of text in an on-line fashion.

RI-Doc

Contrary to sliding window approach used in RI-Word, a RI model built with *document index vectors* first assigns unique index vectors to every document in the training corpus. In the training phase, each word in a document get the respective document vector added to its context vector. The resulting WSM is thus a compressed version of a term-by-document matrix.

RI-ICD

Based on the principle of RI with document index vectors, we here explore a novel way of constructing a WSM by exploiting the ICD-10 code classification done by clinicians. Instead of using document index vectors, we here use *ICD-code index vectors*. First, a unique index vector is assigned to each chapter and sub-chapter in the ICD-10 taxonomy. This means assigning a unique index vector to each "node" in the ICD-10 taxonomy, as illustrated in Figure 2. For each clinical note in the training corpus, the index vector of the their primary ICD-10 code is added to all words within it. In addition, all the index vectors for the ICD-codes higher in the taxonomy are added, each weighted according to their position in the hierarchy. A weight of 1 is given to the full code, while the weight is halved for each step upwards in the hierarchy. The motivation for the latter is to capture a certain degree of similarity between codes that share an initial path in the taxonomy. As a result,

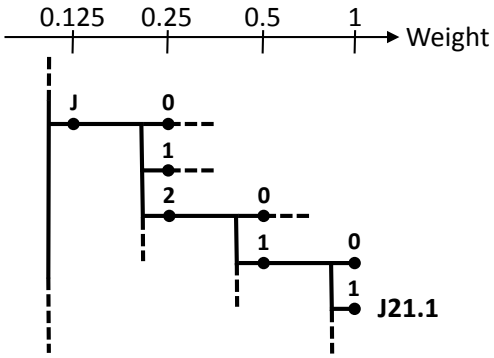

Figure 2: Weighting applied to ICD-code index vectors when training WSMs based on ICD-10 codes (RI-ICD).

this similarity is encoded in the resulting WSM. As a example: for a clinical note labelled with the code $J21.1$, we add the following index vectors to the context vectors of all its constituting words: $iv(J) \times 0.125$, $iv(J2) \times 0.25$, $iv(J21) \times 0.5$ and $iv(J21.1) \times 1.0$. The underlying hypothesis for building a WSM in this way is that it may capture relations between words in a way that better reflects the clinical domain, compared to the other domain-independent methods for constructing a WSM.

RI-Index

As an alternative to using word's (semantic) context vectors, we simply only use their index vectors as their "contextual meaning". When constructing document vectors directly from word index vectors (see Section 4.2), the resulting document vectors represent a compressed version of a document-by-term matrix.

Word2vec

Recently, a novel method for inducing WSMs was introduced by Mikolov et al. (2013a), stemming from the research in deep learning and neural network language models. While the overall objective of learning a continuous vector space representation for each word based on its textual context remains, the underlying algorithms are substantially different from traditional methods such as Latent Semantic Analysis and RI. Considering, in turn, every word in the training data as a target word, the method induces the representations by training a simplified neural network to predict the nearby context words of each target word (skip-gram architecture), or alternatively the target word based on all words in its immediate context (BoW architecture). The vector space representation is subsequently extracted from the learned weights within the neural network. One of the main practical advantages of the word2vec method lies in its scalability, allowing quick training on large amounts of text, setting it apart from the majority of other methods of distributional semantics. Additionally, the word2vec method has been shown to produce representations that surpass in quality traditional methods such as Latent Semantic Analysis, especially on tasks measuring the preservation of important linguistic regularities (Mikolov et al., 2013b).

4.2 Computing care episode similarity

After having computed a WSM, the next step is to build episode vectors to use for the actual retrieval task. This is done by first normalizing the word vectors and multiplying them with a word's TF*IDF weight. An episode vector is then obtained by summing the word vectors of all its words and dividing the result by the total number of words in the episode. Similarity between episodes is determined by computing the cosine similarity between their vectors.

4.3 Baselines

Two baselines were used in this study. The first one is random retrieval of care episodes, which can be expected to give very low scores and serves merely as a sanity check. The second one is Apache Lucene (Cutting, 1999), a state-of-the-art search engine based on look-up of similar documents through a reverse index and relevance ranking based on a TF*IDF-weighted vector space model. Care episodes were indexed using Lucene. Similar to the other models/methods, all of the free text in the query episode, excluding the discharge summary, served as the query string provided to Lucene. Being a state-of-the-art IR system, the scores achieved by Lucene in these experiments should indicate the difficulty of the task.

5 Experiments

In these experiments we strove to have a setup that was as comparable as possible for all models and systems, both in terms of text pre-processing and in terms of the target model dimensionality when inducing the vector space models. The clin-

ical notes are split into sentences, tokenized, and lemmatized using a Constraint-Grammar based morphological analyzer and tagger extended with clinical vocabulary (Karlsson, 1995). After stop words were removed[1], the total training corpus contained 39 million words (minus the query episodes), while the evaluation subset contained 18.5 million words. The vocabulary consisted of 0.6 million unique terms. Twenty care episodes were randomly selected to serve as the query episodes during testing, with the requirement that each had different ICD-10 codes and consisted of a minimum of six clinical notes. The average number of words per query episode is 830.

RI-based and word2vec models have a predefined dimensionality of 800. For RI-based models, 4 non-zeros were used in the index vectors. For the RI-Word model, a narrow context window was employed (5 left + 5 right), weighting index vectors according to their distance to the target word ($weight_i = 2^{1-dist_{it}}$). In addition, the index vectors were shifted once left or right depending on what side of the target word they were located, similar to *direction vectors* as described in (Sahlgren et al., 2008) These parameters for RI were chosen based on previous work on semantic textual similarity (Moen et al., 2013). Also a much larger window of 20+20 was tested, but without noteworthy improvements. The word2vec model is trained with the BoW architecture and otherwise default parameters. In addition to Apache Lucene (version 4.2.0)[2], the word2vec tool[3] was used to train the word2vec model, and the RI-based methods utilized the JavaSDM package[4]. Scores were calculated using the *trec_eval* tool[5].

5.1 Experiment 1: ICD-10 code overlap

In this experiment retrieved episodes with a primary ICD-10 code identical to that of the query episode were considered to be correct. The number of correct episodes varies between 49 and 1654. The total is 7721, and the average is 386. The high total is mainly due to three query episodes with ICD-10 codes that occur very frequently in the episode collection (896, 1590, and 1654 times). When conducting the experiment all care episodes were retrieved for each of the 20 query episodes.

IR model	MAP	P@10
Lucene	0.1379	0.3000
RI-Word	0.0911	0.2650
RI-Doc	0.1015	0.3300
RI-ICD	**0.3261**	**0.5150**
RI-Index	0.1187	0.3200
Word2vec	0.1768	0.3350
Random	0.0154	0.0200

Table 1: Mean average precision and precision at 10 for retrieval of care episodes with the same primary ICD-10 code as the query episode

Performance was measured in terms of mean average precision (MAP) and precision among the top-10 results (P@10), averaged over all 20 queries, as shown in in Table 1. The best MAP score is achieved by RI-ICD, almost twice that of word2vec, which achieved the second best MAP score, whereas RI-Word performed worst of all. All models score well above the random baseline, whereas RI-ICD outperforms Lucene by a large margin. P@10 scores follow the same ranking. The latter scores are more representative for most use cases where users will only inspect the top-n retrieval results.

5.2 Experiment 2: Discharge summary overlap

In this experiment retrieved episodes with a discharge summary similar to that of the query episode were considered to be correct. Using the discharge summaries of the query episodes, the top 100 care episodes with the most similar discharge summary were selected as the most similar care episodes (disregarding the query episode). This was repeated for each of the methods – i.e. the five different semantic models and Lucene – resulting in six different tests. The top 100 was used rather than a threshold on the similarity score, because otherwise six different thresholds would have to be chosen. This procedure thus resulted in six different test collections, each consisting of 20 query episodes with their corresponding 100 most similar collection episodes.

Subsequently a 6-by-6 experimental design was followed where each retrieval method was tested against each test set construction method. At retrieval time, for each query episode, the system retrieves and ranks 1000 care episodes. It can be expected that when identical methods are used for re-

[1] http://www.nettiapina.fi/finnish-stopword-list/

[2] http://archive.apache.org/dist/lucene/java/

[3] https://code.google.com/p/word2vec/

[4] http://www.nada.kth.se/~xmartin/java/

[5] http://trec.nist.gov/trec_eval/

trieval and test set construction, the resulting bias gives rise to relatively high scores. In contrast, averaging over the scores for all six construction methods is assumed to be a less biased indicator of performance.

Table 2 shows the number of correctly retrieved episodes by the different models, with the maximum being 2000 (20 queries times 100 most similar episodes). This gives an indication of the recall among a 1000 retrieved episodes per query, but without caring about precision or ranking. In general, the numbers are relatively good when the same model is used for both retrieval and construction of the test set (cf. values on the diagonal), although in a couple of cases (e.g. with word2vec) results are better with different models. The RI-ICD model performs best when used for both retrieval and test construction. Looking at the averages, which presumably are less biased indicators, RI-ICD and word2vec seem to have comparable performance, with both of them outperforming Lucene. Other models are less successful, although still much better than the random baseline.

The MAP scores in Table 3 show similar results, although here RI-ICD yields the best average score. Both models RI-ICD and word2vec outperform Lucene. Again the RI-ICD model performs exceptionally well when used for both retrieval and test construction.

Finally Table 4 presents precision for top-10 retrieved care episodes. Here RI-Doc yields the best average scores, while RI-ICD and word2vec both perform slightly worse.

6 Discussion

The goal of the experiments was primarily to determine which distributional semantic models work best for care episode retrieval. The experimental results show that several models outperform Lucene at the care episode retrieval task. This suggests that models of higher order semantics contribute positively to calculating document similarities in the clinical domain, compared with straight forward boolean word matching (cf. RI-Index and Lucene).

The relatively good performance of the RI-ICD model, particularly in Experiment 1, suggests that exploiting structured or encoded information in building semantic models for clinical NLP is a promising direction that calls for further investigation. This approach concurs with the arguments in favor of reuse of existing information sources in Friedman et al. (2013). On the one hand, it may not be surprising that the RI-ICD model is performing well on Experiment 1, given how it induces semantic relations between words occurring in episodes with the same ICD-10 code. On the other hand, being able to accurately retrieve care episodes with similar ICD-10 codes evidently has practical value from a clinical perspective.

The different ranking of models in experiments 1 versus 2 confirms that there is a difference between the two indicators of episode similarity, i.e. similarity in terms of their ICD-10 codes versus similarity with regard to their discharge summaries. In our data a single care episode can potentially span across several hospital wards. A better correlation between the similarity measures is to be expected when narrowing the definition of a care episode to only a single ward. Also, taking into consideration all ICD-10 codes for care episodes – not only the primary one – could potentially improve discrimination among care episodes. This could be useful in two ways: (1) to create more precise test sets of the type used in Experiment 1; (2) to extend RI-ICD models with index vectors also for the secondary ICD-10 codes.

Input to the models for training was limited to the free text in the clinical notes, with the exception of the use of ICD-10 codes in the RI-ICD model. Other sources of information could, and probably should, be utilized in a practical care episode retrieval system applied in a hospital, such as the structured and coded information commonly found in EHR systems. Another potential information source is the internal structure of the care episodes, as episodes containing similar notes in the same sequential order are intuitively more likely to be similar. We tried computing exhaustive pairwise similarities between the individual notes from two episodes and then taking the average of these as a similarity measure for the episodes. However, this did not improve performance on any measure. An alternative approach may be to apply sequence alignment algorithms, as commonly used in bioinformatics (Gusfield, 1997), in order to detect if both episodes contain similar notes in the same temporal order. We leave this to future work.

IR model \ Test set	Lucene	RI-Word	RI-Doc	RI-ICD	RI-Index	Word2vec	Average	Rank
Lucene	889	700	670	687	484	920	725	2
RI-Word	643	800	586	600	384	849	644	5
RI-Doc	665	630	859	697	436	795	680	4
RI-ICD	635	459	659	**1191**	490	813	707	3
RI-Index	690	491	607	654	576	758	629	6
Word2vec	789	703	702	870	516	**1113**	**782**	1
Random	74	83	86	67	84	85	79	7

Table 2: Number of correctly retrieved episodes (max 2000) for different IR models (rows) when using different models for measuring discharge summary similarity (columns)

IR model \ Test set	Lucene	RI-Word	RI-Doc	RI-ICD	RI-Index	Word2vec	Average	Rank
Lucene	0.0856	0.0357	0.0405	0.0578	0.0269	0.0833	0.0550	3
RI-Word	0.0392	0.0492	0.0312	0.0412	0.0151	0.0735	0.0416	6
RI-Doc	0.0493	0.0302	0.0677	0.0610	0.0220	0.0698	0.0500	4
RI-ICD	0.0497	0.0202	0.0416	**0.1704**	0.0261	0.0712	**0.0632**	1
RI-Index	0.0655	0.0230	0.0401	0.0504	0.0399	0.0652	0.0473	5
Word2vec	0.0667	0.0357	0.0404	0.0818	0.0293	0.1193	**0.0622**	2
Random	0.0003	0.0003	0.0005	0.0002	0.0003	0.0004	0.0003	7

Table 3: Mean average precision for different IR models (rows) when using different models for measuring discharge summary similarity (columns)

IR model \ Test set	Lucene	RI-Word	RI-Doc	RI-ICD	RI-Index	Word2vec	Average	Rank
Lucene	0.2450	0.1350	0.1200	0.1650	0.0950	0.1900	0.1583	5
RI-Word	0.1350	0.1500	0.1000	0.1350	0.0600	0.2100	0.1316	6
RI-Doc	0.2000	0.1250	0.2050	0.2200	0.0900	0.2400	**0.1800**	1
RI-ICD	0.1700	0.0650	0.1350	**0.3400**	0.0950	0.2050	0.1683	2
RI-Index	0.2000	0.1250	0.1550	0.1250	0.1700	0.2050	0.1633	3
Word2vec	0.1800	0.1200	0.1150	0.2100	0.0850	0.2650	0.1625	4
Random	0.0000	0.0000	0.0050	0.0000	0.0000	0.0000	0.0008	7

Table 4: Precision at top-10 retrieved episodes for different IR models (rows) when using different models for measuring discharge summary similarity (columns)

7 Conclusion and future work

In this paper we proposed the task of *care episode retrieval* as a way of evaluating several distributional semantic models in their performance at IR. As manually constructing a proper test set of classified care episodes is costly, we experimented with building test sets by exploiting either ICD-10 code overlap or semantic similarity of discharge summaries. A novel method for generating semantic models utilizing the ICD-10 codes of care episodes in the training corpus was presented (RI-ICD). The models, as well as the Lucene search engine, were applied to the care episode retrieval task and their performance was evaluated against the test sets using different evaluation measures. The results suggest that the RI-ICD model is better suited to IR tasks in the clinical domain compared with models trained on local distributions of words, or those relying on direct word matching. The word2vec model performed relatively well and outperformed Lucene in both experiments.

In the results reported here, the internal sequence of clinical notes is ignored. Future work should focus on exploring the temporal (sub-) sequence similarities between care episode pairs for doing care episode retrieval. Further work should also focus on expanding on the RI-ICD method by exploiting other types of structured and/or encoded information related to clinical notes for training semantic models tailored for NLP in the clinical domain.

Acknowledgments

This study was partly supported by the Research Council of Norway through the EviCare project (NFR project no. 193022), the Turku University Hospital (EVO 2014), and the Academy of Finland (project no. 140323). The study is a part of the research projects of the Ikitik consortium (`http://www.ikitik.fi`). We would like to thank Juho Heimonen for assisting us in preprocessing the data and the reviewers for their helpful comments.

References

Helen Allvin, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravičius, Martin Hassel, Dimitrios Kokkinakis, Heljä Lundgren-Laine, Gunnar Nilsson, Øystein Nytrø, et al. 2010. Characteristics and analysis of finnish and swedish clinical intensive care nursing narratives. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 53–60. Association for Computational Linguistics.

Doug Cutting. 1999. Apache Lucene open source package.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35(4):222–235.

Carol Friedman, Thomas C Rindflesch, and Milton Corn. 2013. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765–773.

Dan Gusfield. 1997. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravi, Martin Duneld, et al. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of biomedical semantics*, 5(1):6.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Fred Karlsson. 1995. *Constraint grammar: a language-independent system for parsing unrestricted text*. Mouton de Gruyter, Berlin and New York.

Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2439–2442. ACM.

Mario Lenz, André Hübner, and Mirjam Kunze. 1998. Textual cbr. In *Case-based reasoning technology*, pages 115–137. Springer.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751. Association for Computational Linguistics, June.

Hans Moen, Erwin Marsi, and Björn Gambäck. 2013. Towards dynamic word sense discrimination with random indexing. *ACL 2013*, page 83.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

Alan L Rector. 1999. Clinical terminology: why is it so hard? *Methods of information in medicine*, 38(4/5):239–252.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Hagit Shatkay. 2005. Hairpins in bookstacks: information retrieval from biomedical text. *Briefings in Bioinformatics*, 6(3):222–238.

Johanna I Westbrook, Enrico W Coiera, and A Sophie Gosling. 2005. Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*, 12(3):315–321.

World Health Organization and others. 2013. International classification of diseases (icd).

Lei Yang, Qiaozhu Mei, Kai Zheng, and David A Hanauer. 2011. Query log analysis of an electronic health record search engine. In *AMIA Annual Symposium Proceedings*, volume 2011, page 915. American Medical Informatics Association.