

Text2Node: a Cross-Domain System for Mapping Arbitrary Phrases to a Taxonomy

Rohollah Soltani and Alexandre Tomberg

Knowtions Research Inc.
Toronto, Ontario, Canada

Abstract

Electronic health record (EHR) systems are used extensively throughout the healthcare domain. However, data interchangeability between EHR systems is limited due to the use of different coding standards across systems. Existing methods of mapping coding standards based on manual human experts mapping, dictionary mapping, symbolic NLP and classification are unscalable and cannot accommodate large scale EHR datasets.

In this work, we present Text2Node, a cross-domain mapping system capable of mapping medical phrases to concepts in a large taxonomy (such as SNOMED CT). The system is designed to generalize from a limited set of training samples and map phrases to elements of the taxonomy that are not covered by training data. As a result, our system is scalable, robust to wording variants between coding systems and can output highly relevant concepts when no exact concept exists in the target taxonomy. Text2Node operates in three main stages: first, the lexicon is mapped to word embeddings; second, the taxonomy is vectorized using node embeddings; and finally, the mapping function is trained to connect the two embedding spaces. We compared multiple algorithms and architectures for each stage of the training, including GloVe and FastText word embeddings, CNN and Bi-LSTM mapping functions, and node2vec for node embeddings. We confirmed the robustness and generalisation properties of Text2Node by mapping ICD-9-CM Diagnosis phrases to SNOMED CT and by zero-shot training at comparable accuracy.

This system is a novel methodological contribution to the task of normalizing and linking phrases to a taxonomy, advancing data interchangeability in healthcare. When applied, the system can use electronic health records to generate an embedding that incorporates taxonomical medical knowledge to improve clinical predictive models.

1 Introduction

Electronic health and medical record (EHR/EMR) systems are steadily gaining in popularity. Ever more facets of healthcare are recorded and coded in such systems, including patient demographics, disease history and progression, laboratory test results,

clinical procedures and medications, and even genetics. This trove of information is a unique opportunity to learn patterns that can help improve various aspects of health-care. However, the sheer number of various coding systems used to encode this clinical information is a major challenge for anyone trying to analyse structured EHR data. Even the most widely used coding systems have multiple versions to cater to different regions of the world. Software developed to analyze one version of the coding system may not be used for another version, let alone a different coding system. In addition to the public and well documented coding systems, a multitude of private coding mechanisms that have no mappings to any public coding systems are used by insurance companies and certain hospitals.

The efforts to solve this problem range from the development of mapping dictionaries between coding systems to machine learning driven approaches. One example of the former is cTAKES [27], a project that uses annotated lookup tables to map clinical entities to concepts in a controlled vocabulary such as SNOMED CT (Systematised Nomenclature of Medicine – Clinical Terms, [20]). The reliance of cTAKES and similar systems on symbolic natural language processing techniques makes them hard to generalise and scale, especially in view of regular updates and changes to the target vocabulary. An example of the latter approach is the work of [39], where word embeddings have been used in a hierarchical structure to annotate and map medical concepts to a reference taxonomy. Their method is based on classification learning and limited in its ability to be applied on controlled vocabularies such as SNOMED CT due to the large size of these vocabularies and small number of terms or phrases associated with each concept.

In this work, we present a system that we call Text2Node that maps phrases from a medical lexicon to a concept in a large target taxonomy (we use SNOMED CT in this paper). Our system is scalable since it relies on lexicon and taxonomy embeddings that are generated in an unsupervised fashion. It is robust to changes in the phrase vocabulary, can output similar concepts in cases when the exact match is not found, and can map to parts of the taxonomy that were not covered by the training data. We achieved this result by relying on node embeddings built from the relationships between the concepts in SNOMED CT, and by replacing the classification problem of finding the best concept with a regression to the nearest node embedding.

Applications of the Text2Node system include scalable generation of mapping dictionaries for non-public medical coding systems, identification and recognition of clinical concepts or named entities in EHRs and EMRs, and generation of phrases representations that are medical concept aware.

Technical Significance We developed Text2Node as a system that maps phrases from a medical lexicon to concepts in a large target taxonomy. The system relies on word and node embeddings that are generated in a scalable unsupervised fashion; it is robust to changes in the vocabulary and presence of out-of-vocabulary words; it returns highly relevant concepts even in cases where the exact match is not found or does not exist; and it generalises well as demonstrated by our zero-shot learning experiments. A key ingredient in achieving these results was the transformation of the classification task of finding the best concept into a regression to the nearest node

embedding problem.

Clinical Relevance Our Text2Node system can be used for generating mapping dictionaries for non-public medical coding systems to large public taxonomies in a scalable way. It is robust to changes in wording between coding systems, it can generalise well from a limited set of training samples, and it can map to elements of taxonomies that are not covered by training data. The Text2Node system can also be used for identification and recognition of clinical concepts and clinical named entities in electronic medical records.

2 Motivation

A natural approach to mapping an arbitrary coding system to a well documented public taxonomy is to start with the phrases that describe each code. Since the coding system must be understandable to its human users, the phrases describing each code must provide enough information for them to be able to use it.

Lexical representation Since the seminal work of [24] on Word2Vec, word embeddings have been widely used to capture the *semantic meaning* of words, phrases and even sentences. For example, word embeddings have been successfully applied in clinical settings to information retrieval, named entity recognition and patient outcome prediction tasks on unstructured text in EMR narratives. Specifically, this technique assigns real-valued vectors of a fixed length to individual words from a collection of documents, called a *corpus*. This vector representation is such that it captures the semantic relationships between words, so that synonyms lie close to each other, while unrelated words are far away. Following Word2Vec, many algorithms have been developed and used successfully to generate word embeddings, including GloVe [26] and FastText [4].

The defining feature of all of word embedding algorithms is the use of contextual interchangeability as a proxy for relatedness in meaning. However, this feature can be a problem for the task at hand, especially in the medical setting. For example, the terms *cold* and *viral respiratory infection* are highly related, but are not often used interchangeably in the medical context. The use of contextual interchangeability as a proxy will lead the word algorithms to incorrectly position the vectors corresponding to these two terms very far from each other in the embedding space.

Taxonomical representation In healthcare and biomedical research, the relationships between entities contain valuable information, because they describe the interactions and causal relationships between diagnosis, medications and procedures, as well as genetic components. To document the complex relationships, large databases have been built, including biomedical knowledge graphs (e.g. PharmGKB [16], DrugBank [37]), ontologies (e.g. Gene Ontology [9]) and taxonomies such as International Statistical Classification of Diseases (ICD), and SNOMED CT.

Network topology is used to analyse and represent the network structure of these biomedical databases. Such analysis requires high computational costs due to the

high dimensionality and sparsity of these databases. Network embedding technologies provide new effective paradigms to solve the network analysis problem. It converts network into a low-dimensional space while maximally preserving its structural properties. Many network embedding algorithms have been developed to embed these graphs into vector spaces (see for example [29]) and used successfully to predict drug-drug interactions [34] and another paper.

Mapping between representations Since contextual interchangeability is not a good metric of medical relatedness, word embeddings cannot be directly used to map between coding systems. However, they can be used to capture semantic information from phrases that are used to describe such systems. In contrast, node embeddings generated from concepts in a medical taxonomy are a better representation of medical relatedness, because they are built from relationships between medical concepts. In order to bridge the gap between these two embedding spaces, a mapping function is needed. This mapping function is to operate on the level of vector representations rather than original phrases and concepts. This has two important benefits: these vector spaces are low-dimensional compared to hundreds of thousands of original concepts; the function learned from embeddings should be more generalisable and is in general easier to train.

Zero-shot learning Whenever there is scarcity of supervised data, machine learning models often fail to carry out reliable generalisations. Obtaining correctly labelled data is often costly and impractical for large datasets. A practical application of concept embedding is the zero-shot transformation of words and concepts that were missing in the training data. It is possible to generalize the mapping function and accurately map unseen concepts, despite having only a few training examples per concept, because embedding training in both domains is an unsupervised task. This is done through nearest neighbour retrieval, where the closest embedding in the target space is selected according to some similarity metric.

3 Methods

A visual representation of the Text2Node system is provided by Figure 1. There are three main stages: first, the lexicon is mapped to word embeddings; second, the taxonomy is vectorized using node embeddings; and finally, the mapping function is trained to connect the two embedding spaces.

3.1 Word Embeddings

Corpus According to [36], word embeddings trained on a biomedical corpus can capture the semantic meaning of medical concepts better than embeddings trained on an unspecialised set of documents. We have thus constructed a large corpus consisting of open access papers from PubMed, free text admission and discharge notes from the MIMIC-III Clinical Database [14, 12], narratives from the US Food and Drug

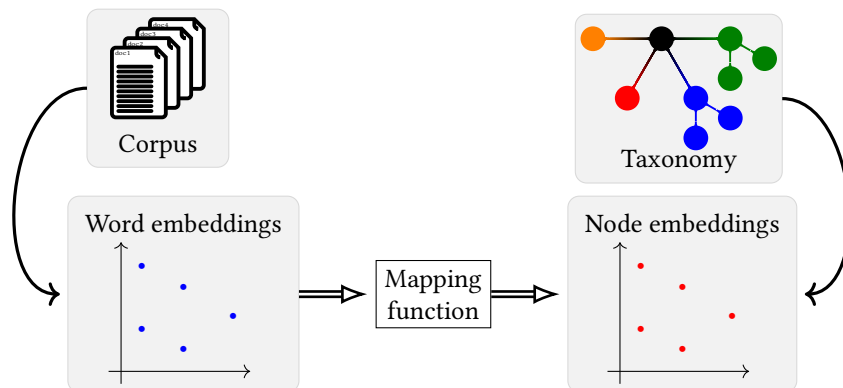


Figure 1: Text2Node System overview

Administration (FDA) Adverse Event Reporting System (FAERS), and a part of the 2010 Relations Challenge from i2b2 [33].

The documents from those sources were pre-processed using the Stanford CoreNLP pipeline [23] to split sentences, add spaces around punctuation marks, change all characters to lowercase, and reformat to one sentence per line. Finally, all files were concatenated into a single document with 235M sentences and 6.25B words. This file was used to train both algorithms described below.

Algorithms We focused on two algorithms for learning word embeddings: GloVe and FastText. An important distinction between them is the treatment of words that are not part of the training vocabulary: GloVe creates a special out-of-vocabulary token and maps all of these words to this token’s vector, while FastText uses subword information to generate an appropriate embedding.

We set the vector space dimensionality to 200 and the minimal number of word occurrences to 10 for both algorithms: this produced a vocabulary of 3.6M tokens. All other algorithm-specific settings were left at their default values, since we expected that Text2Node would perform just as well with different versions of embeddings.

3.2 Node Embeddings

Taxonomy The taxonomy to which Text2Node maps phrases can be arbitrary. In this paper, we focused on the 2018 international version of SNOMED CT as our target graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertex set \mathcal{V} consists of 392K medical concepts and the edge set \mathcal{E} is composed of 1.9M relations between the vertices including `is_a` relationships and attributes such as `finding_site` and `due_to`.

Algorithm To construct taxonomy embeddings, we used node2vec method from [13]. We started a random walk on the edges from each vertex $v \in \mathcal{V}$ and stopped it after a fixed number of steps (20 in our application). All the vertices visited by the walk were considered to be part of the graph neighbourhood $N(v)$ of v . Following the

skip-gram architecture from [24], we selected the feature vector assignment function $v \mapsto f_{n2v}(v) \in \mathbb{R}^{128}$ by solving the optimisation problem

$$f_{n2v} = \operatorname{argmax}_f \sum_{u \in \mathcal{V}} \log \mathbb{P}[N(u)|f(u)],$$

using stochastic gradient descent and negative sampling.

3.3 Mapping

The mapping between phrases and concepts in the target taxonomy was made by associating points in the node embedding vector space to sequences of word embeddings corresponding to individual words in a phrase. As illustrated in Figure 2, given a phrase consisting of n words with the associated word embeddings w_1, \dots, w_n , the mapping function was $m : (w_1, \dots, w_n) \mapsto p$, where $p \in \mathbb{R}^{128}$ is a point in the node embedding vector space. To complete the mapping, we found concepts in the

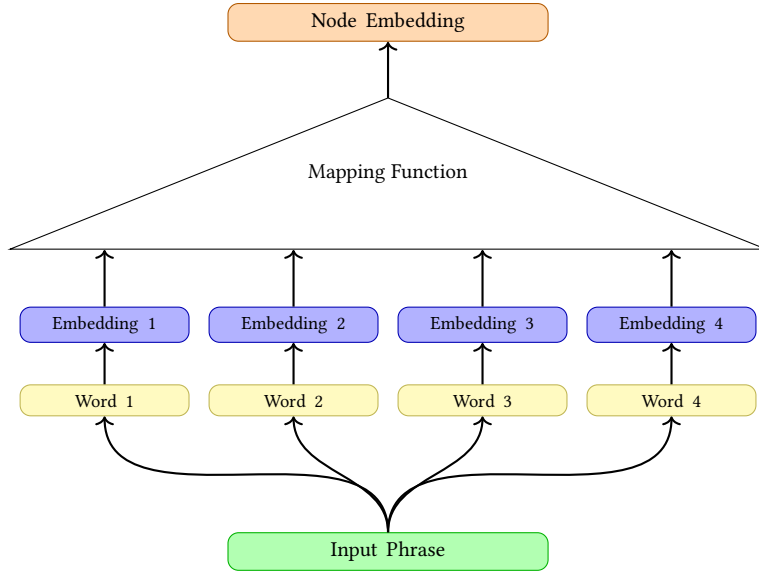


Figure 2: Mapping procedure: the input phrase is split into words that are converted to word embeddings and fed into the mapping function; the output of the function is a point in the node embedding space \mathbb{R}^{128} .

taxonomy whose node embeddings were the closest to the point p . We experimented with two measures of closeness in the node embedding vector space \mathbb{R}^{128} : Euclidean ℓ_2 distance and cosine similarity, that is

$$\begin{aligned} \ell_2 \text{ distance } (p, q) &= \|p - q\| = \sqrt{(p - q) \cdot (p - q)}, \\ \cos \text{ similarity } (p, q) &= \frac{p \cdot q}{\|p\| \|q\|}. \end{aligned}$$

In some cases, for example to compute the top- k accuracy of the mapping (see Section 4), a list of k closest concepts was used.

Function architectures The exact form of the mapping function m can vary. In this paper, we tested three architectures: a linear mapping, a convolutional neural network (CNN), and a bidirectional long short term memory network (Bi-LSTM). We padded or truncated all phrases to be exactly 20 words long to represent each phrase by 20 word embeddings $w_1, \dots, w_{20} \in \mathbb{R}^{200}$ in order to accommodate all three architectures.

Linear We concatenated 20 word embeddings into a single 4000 dimensional vector w , and the linear mapping was given by $p = m(w) = Mw$ for a 4000×128 matrix M .

CNN A convolutional neural network [17] applies convolutional filters of different sizes to the input vectors. The feature maps produced by the filters are then fed into a pooling layer followed by a projection layer to obtain an output of desired dimension. Here, we used filters representing word windows of sizes 1, 2, 3, and 5, followed by a maximum pooling layer and a projection layer to 128 output dimensions. CNN is a non-linear transformation that is more flexible and may be better suited to capture complex patterns in the input than a linear mapping method. Another useful property of the CNN is its ability to learn invariant features regardless of their position in the phrase.

Bi-LSTM A bidirectional long short term memory network (Bi-LSTM) [28] is also a non-linear transformation. This type of neural network operates by recursively applying a computation to every element of the input sequence conditioned on the previous computed results in both forward and backward directions. Bi-LSTM excels at learning long distance dependencies in its input. We built a Bi-LSTM to approximate the mapping function m using a single Bi-LSTM cell with 200 hidden units followed by a projection layer to 128 output dimensions.

Training We gathered training data consisting of phrase-concept pairs from the taxonomy itself. Because most nodes in SNOMED CT have multiple phrases describing them (synonyms), we considered each synonym-concept pair separately for a total of 269K training examples. To find the best mapping function m_* in each of the three architectures described above, we solved the supervised regression problem

$$m_* = \operatorname{argmin}_m \sum_{(\text{phrase}, \text{node})} \|m(\text{phrase}) - f_{\text{n2v}}(\text{node})\|_{\ell_2}^2,$$

using the Adam optimizer [15] for 50 epochs.

4 Evaluation and Results

First, we evaluated all mapping function architectures to determine the top performing ones using a random hold-out test set of 10K phrase-concept pairs. Next, we

tested the robustness and generalisability of the winning architectures on a new lexicon consisting of 8.9K phrases from ICD-9-CM. Finally, we tested the zero-shot learning capability of Text2Node. We did this by randomly selecting 1K concepts from our training set, removing all associated 3.4K phrase–concept pairs from the training dataset, and retraining the system. The performance was evaluated by mapping the removed phrases to concepts that were never seen before.

In these tests we considered two performance metrics: accuracy and mean graph distance.

Accuracy is the proportion of test phrases that are mapped exactly to their corresponding concept (higher is better). This metric is often used for classification based approaches and is a standard metric for the mapping problem.

Mean graph distance is the average graph distance (length of the shortest path) in the taxonomy between the target concept and the predicted concept (lower is better). This metric tells us *how far* is the predicted concept from the exact match, and low scores are evidence for predictions that are immediate neighbours of the target concept, that is its children or parents.

Due to the probabilistic nature of node embeddings and the topology of our target graph, it was possible that two distinct concepts were embedded to the exact same location in the node embedding space. Since the mapping procedure involved nearest neighbour searches, we computed the performance metrics for the top- k results of these searches for $k = 1, 5, 10, 20, 50$. This also allowed us to better cover the test cases where multiple concepts were been assigned to a phrase.

4.1 Intrinsic Evaluation

We used a random hold-out test set of 10K phrase–concept pairs to find the top performing architectures. The accuracy results are in Table 1, and the graph distance results are in Table 2. The overall best system was the Bi-LSTM mapping function on top of FastText embeddings. An interesting finding was that the cosine similarity yields consistently better accuracy scores, while the ℓ_2 distance provides for slightly smaller average graph distances.

The Bi-LSTM on FastText architecture achieved 23.9% top-1 accuracy with the mean graph distance of 1.27. This confirmed our expectation that the majority of cases that were not mapped to the exact target concept were still mapped to one of its immediate neighbours. The accuracy increased to 67.1% when we considered the top-20 closest matches.

4.2 Extrinsic Evaluation (ICD-9-CM)

To check the robustness and generalisability of the Text2Node system, we created an extrinsic evaluation task consisting of 8.9K ICD-9-CM phrases mapped by medical experts from [32] to a unique SNOMED CT concept. The lexicon of these phrases is different from our target taxonomy, because ICD-9-CM was developed for assigning codes to diagnoses and procedures associated with hospital utilization in the United

Text2Node model			Accuracy (top k)				
Mapping	Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
Linear	GloVe	cos	0.033	0.106	0.163	0.231	0.352
Linear	GloVe	ℓ_2	0.014	0.062	0.099	0.153	0.245
Linear	FastText	cos	0.029	0.094	0.153	0.226	0.355
Linear	FastText	ℓ_2	0.012	0.050	0.091	0.143	0.237
CNN	GloVe	cos	0.061	0.176	0.249	0.338	0.485
CNN	GloVe	ℓ_2	0.045	0.150	0.214	0.294	0.419
CNN	FastText	cos	0.082	0.218	0.304	0.397	0.546
CNN	FastText	ℓ_2	0.067	0.181	0.256	0.350	0.495
Bi-LSTM	GloVe	cos	0.225	0.442	0.543	0.640	0.762
Bi-LSTM	GloVe	ℓ_2	0.195	0.402	0.497	0.602	0.722
Bi-LSTM	FastText	cos	0.239	0.466	0.571	0.671	0.785
Bi-LSTM	FastText	ℓ_2	0.201	0.416	0.515	0.616	0.744

Table 1: Intrinsic evaluation results: accuracy on a random hold-out test set of 10K phrase-concept pairs.

Text2Node model			Mean graph distance (top k)				
Mapping	Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
Linear	GloVe	cos	2.08	2.23	2.30	2.35	2.42
Linear	GloVe	ℓ_2	1.79	2.07	2.17	2.24	2.32
Linear	FastText	cos	2.01	2.17	2.23	2.28	2.35
Linear	FastText	ℓ_2	1.78	2.06	2.16	2.23	2.31
CNN	GloVe	cos	1.71	1.95	2.03	2.11	2.19
CNN	GloVe	ℓ_2	1.60	1.88	1.98	2.06	2.14
CNN	FastText	cos	1.60	1.86	1.95	2.03	2.13
CNN	FastText	ℓ_2	1.52	1.80	1.91	2.00	2.09
Bi-LSTM	GloVe	cos	1.28	1.63	1.76	1.88	2.01
Bi-LSTM	GloVe	ℓ_2	1.26	1.61	1.74	1.86	1.98
Bi-LSTM	FastText	cos	1.27	1.61	1.75	1.87	2.01
Bi-LSTM	FastText	ℓ_2	1.25	1.59	1.73	1.85	1.98

Table 2: Intrinsic evaluation results: mean graph distance on a random hold-out test set of 10K phrase-concept pairs.

States. The accuracy results are in Table 3, and the graph distance results are in Table 4.

Text2Node model		Accuracy (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	0.202	0.430	0.541	0.644	0.755
GloVe	ℓ_2	0.170	0.394	0.501	0.602	0.714
FastText	cos	0.210	0.443	0.553	0.658	0.763
FastText	ℓ_2	0.181	0.403	0.507	0.614	0.725

Table 3: Extrinsic evaluation: accuracy of the Bi-LSTM mapping function on 8.9K ICD-9-CM phrases.

Text2Node model		Mean graph distance (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	1.42	1.74	1.88	2.00	2.14
GloVe	ℓ_2	1.40	1.72	1.85	1.97	2.10
FastText	cos	1.40	1.73	1.87	1.99	2.14
FastText	ℓ_2	1.36	1.69	1.84	1.95	2.09

Table 4: Extrinsic evaluation: mean graph distance of the Bi-LSTM mapping function on 8.9K ICD-9-CM phrases.

The Bi-LSTM model on FastText also yielded the best score in this experiment. When using cosine similarity as similarity measurement, it achieved 21% in top-1 accuracy and 44% in top-5 accuracy, and when ℓ_2 metric was used, it achieved 1.36 in graph distance. The graph distance of 1.36 shows that, although the exact correct match was only 21% at top-1 of Text2Node ranking, all the predicted concepts were close to the exact match in terms of taxonomy distance, that is they are either synonyms, parents, or children of the exact match concept from 392K nodes.

This test set also allowed us to study the effect of extra knowledge of source phrases on mapping to the target taxonomy. In our previous evaluations, we mapped phrases to one of the 392K concepts in the target taxonomy. However, when we reduced the search space to 7.5K concepts that have at least one ICD-9-CM phrase mapped to them, the accuracy and mean graph distance results improved significantly as shown in Tables 5 and 6.

Text2Node model		Accuracy (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	0.379	0.698	0.793	0.855	0.905
GloVe	ℓ_2	0.350	0.655	0.754	0.822	0.876
FastText	cos	0.389	0.710	0.804	0.863	0.914
FastText	ℓ_2	0.362	0.658	0.760	0.829	0.887

Table 5: Extrinsic evaluation: accuracy of the Bi-LSTM mapping function on 8.9K ICD-9 phrases with the search space reduced to 7.5K concepts.

Text2Node model		Mean graph distance (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	1.32	1.80	1.98	2.13	2.29
GloVe	ℓ_2	1.36	1.82	1.98	2.12	2.26
FastText	cos	1.27	1.78	1.98	2.13	2.28
FastText	ℓ_2	1.31	1.80	1.97	2.11	2.25

Table 6: Extrinsic evaluation: mean graph distance of the Bi-LSTM mapping function on 8.9K ICD-9 phrases with the search space reduced to 7.5K concepts.

SNOMED CT consists of a wide category of medical concepts including diseases, procedures, medications, and body structures that are closely related, meaning that these nodes are close in the node embedding space. Therefore, extra knowledge about the category of the taxonomy allowed us to reduce the search space and boost the performance of Text2Node. In this test, the accuracy for top-1 almost doubled to 39% and the graph distance went down to 1.27 by filtering the output concepts to be correlated to the ICD-9 concepts.

4.3 Zero-shot Evaluation

To evaluate the zero-shot learning capability, we randomly selected 1K concepts from the taxonomy that appeared in our training set. We then removed all 3.4K phrase-concept pairs associated to the selected concepts from the training set, and used them as the test set. We retrained the Bi-LSTM mapping network from scratch using the new training set, so that all of the targets in the zero-shot test set were never seen before by the mapping function. The accuracy and mean graph distance results are given in Tables 7 and 8.

Text2Node model		Accuracy (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	0.242	0.449	0.538	0.622	0.737
GloVe	ℓ_2	0.227	0.391	0.481	0.583	0.706
FastText	cos	0.221	0.499	0.591	0.671	0.780
FastText	ℓ_2	0.186	0.439	0.534	0.636	0.734

Table 7: Zero-shot evaluation: accuracy of a newly trained Bi-LSTM mapping function on 3.4K phrases mapping to 1K previously unseen concepts.

In this test, we were able to achieve comparable results to the intrinsic task evaluations from Section 4.1, and we showed that our system is general enough to be able to map to unseen concepts in the training set.

Text2Node model		Mean graph distance (top k)				
Word Emb	Metric	$k = 1$	$k = 5$	$k = 10$	$k = 20$	$k = 50$
GloVe	cos	1.41	1.74	1.86	1.97	2.09
GloVe	ℓ_2	1.38	1.70	1.82	1.93	2.04
FastText	cos	1.41	1.69	1.81	1.92	2.05
FastText	ℓ_2	1.38	1.65	1.78	1.88	2.00

Table 8: Zero-shot evaluation: mean graph distance of a newly trained Bi-LSTM mapping function on 3.4K phrases mapping to 1K previously unseen concepts.

5 Discussion and Related Work

5.1 Discussion

In this paper we have described our Text2Node system for mapping arbitrary medical phrases to a large taxonomy. In our system, we first used embedding models to project phrases in the source medical lexicon domain and concept entities in the target taxonomy domain into two different latent spaces. We then learned a mapping function between those latent spaces to associate points in the node embedding space to sequences of word embedding vectors. We used nearest neighbour search in the target domain to find candidate medical concepts for a given phrase. We tested a number of word embedding algorithms and mapping function architectures and we obtained the best result using a Bi-LSTM mapping model and FastText word embeddings.

In summary, our experiments showed that Text2Node is a robust and generalisable system that achieved a satisfactory result at mapping ICD-9-CM Diagnosis phrases to SNOMED CT and a comparable accuracy at zero-shot training. More generally, we confirmed our hypothesis that utilising the hierarchical structure of a biomedical taxonomy to learn node representations can achieve a better mapping system between medical phrases and taxonomy concepts.

5.2 Related work

Concept mapping is a longstanding problem and there is a variety of existing approaches. For example, cTAKES [27], MetaMap [1], MedLEE [11], KnowledgeMap [10] NOBLE [31], and PDD Graph [35] use annotated lookup tables in order to map clinical entities to concepts in controlled vocabularies such as SNOMED CT and the Unified Medical Language System (UMLS). To do so, they rely on symbolic natural language processing techniques, which makes them hard to generalise and scale. cTAKES, was developed by Mayo Clinic and later transitioned to an Apache project. MetaMap was developed by the National Library of Medicine for mapping biomedical text to the UMLS Metathesaurus. MedLEE is one of the earliest clinical NLP systems developed and is mostly used for pharmacovigilance.

Word embeddings have been widely used in medical domain since they contain rich latent semantic information. A few prominent examples include [25] that trained a set of word embeddings on PubMed data using Word2Vec; [19] that studied drug name recognition utilizing word embeddings; [18] that applied word embedding to

extract events from the biomedical text; and [38] that examined the use of word embeddings for clinical abbreviation disambiguation.

Recently researchers have started to explore the possibility of efficient representation learning in the medical domain. [6] utilised a hierarchical structure from the medical term ontology by applying an attention mechanism. [7, 5, 8] learned the representation of clinical codes using skip-gram from structured visit records of patients. [30] used a modified restricted RBM for representing EHR concepts using a structured training process. In a similar paper, [21] used AEs to generate concept vectors from word-based concepts extracted from clinical free text.

For cross-domain mapping in natural language processing, [2] used linear transformation to improve bilingual word embedding mapping, and [3] used adversarial training for cross-lingual word embeddings without parallel data. Finally, in recommendation systems, [22] used a cross-domain recommendation system to leverage feedbacks or ratings from multiple domains and improve recommendation performance.

5.3 Text2Node Applications

NER The problem of named entity recognition (NER) is the task of extracting relevant concepts from free text. Once extracted, such concepts need to be mapped to a taxonomy of known entities. Our Text2Node system can be used to solve this mapping problem.

Medical phrase representation Finding a representation for a phrase that can capture all the information in its sequence of words is a challenging task in natural language processing. The mapping function in our Text2Node system can be seen as an encoding model that is generating a representations for medical phrases. Since these representations have been trained in a supervised manner using the information from the taxonomy, they also represent their taxonomy position and the structure of their neighbourhood in the taxonomy. Thus this supervised representation of the medical phrases can be used in different tasks (i.e. medical outcome prediction) using transfer learning techniques.

Zero-shot learning A practical application of embedding mappings is the zero-shot transformation of concepts that were missing in the training data. Our system showed a satisfactory performance therefore it can be used for data pre-processing when the training data is hard to obtain as is often the case.

5.4 Future work

Our system showed a satisfactory performance in mapping phrases between two taxonomies and in a promising result towards solving the zero-shot problem. More work can be done to make it better.

In EHRs, there are often phrases with a complex internal structure. For example, a surgery code labelled *Eye, Intra-ocular Foreign Body, Removal from Anterior Segment* could be decomposed into three parts: location, problem, operation. The performance

of our system is yet to be evaluated by focusing on solely complex phrases. On the other hand, some NLP features such as part of speech can be coupled into our system to solve this problem.

In the taxonomy, there exist nodes that have the exact same connections to the rest of the taxonomy. Therefore, they have the same representation in our node embedding space. Our current system does not address this issue and these nodes will not be differentiated. This problem can be solved by adding extra informative connections to those nodes to make their neighbourhoods different. It is also possible to use two taxonomies to train the node embeddings, to differentiate these nodes and yield better performance.

References

- [1] Alan R. Aronson and Francois-Michel Lang. An overview of metamap: historical perspective and recent advances. *JAMIA*, 17(3):229–236, 2010.
- [2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [5] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.
- [6] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.
- [7] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction, 2016, arXiv preprint arXiv:1602.03686.
- [8] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.
- [9] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 11 2014, <http://oup.prod.sis.lan/nar/article-pdf/43/D1/D1049/7329240/gku1179.pdf>.
- [10] Joshua C. Denny, Plomarz R. Irani, Firas H. Wehbe, Jeffrey D. Smithers, and Anderson Spickard III. The knowledgemap project: Development of a concept-based medical school curriculum database. In *AMIA*. AMIA, 2003.
- [11] Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 03 1994.
- [12] Ary L Goldberger et al. Components of a new research resource for complex physiologic signals, physiobank, physiotoolkit, and physionet, american heart association journals. *Circulation*, 101(23):1–9, 2000.
- [13] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

- [14] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
- [16] Teri E Klein, Jeffrey T Chang, Mildred K Cho, Katrina L Easton, Ray Ferguson, Micheal Hewett, Zhen Lin, Y Liu, S Liu, DE Oliver, et al. Integrating genotype and phenotype information: an overview of the pharmgkb project. *The pharmacogenomics journal*, 1(3):167, 2001.
- [17] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [18] Chen Li, Runqing Song, Maria Liakata, Andreas Vlachos, Stephanie Seneff, and Xiangrong Zhang. Using word embedding for bio-event extraction. In *Proceedings of BioNLP 15*, pages 121–126, 2015.
- [19] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information*, 6(4):848–865, 2015.
- [20] Esther Lozano, Mar Marcos, Begoa Martinez-Salvador, Albert Alonso, and Josep Ramon Alonso. Experiences in the development of electronic care plans for the management of comorbidities. In *Lecture Notes in Computer Science*, volume 5943, pages 113–123, Berlin, Heidelberg, 07 2010. Springer.
- [21] Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. Clinical relation extraction with deep learning. *IJHIT*, 9(7):237–248, 2016.
- [22] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. Cross-domain recommendation: An embedding and mapping approach. In *IJCAI*, pages 2464–2470, 2017.
- [23] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.
- [24] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
- [25] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44, 2013.
- [26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [27] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
- [28] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [29] Chang Su, Fei Wang, Jie Tong, Yongjun Zhu, and Peng Cui. Network embedding in biomedical data science, 12 2018, <https://doi.org/10.1093/bib/bby117>.
- [30] Truyen Tran, Tu Dinh Nguyen, Dinh Q. Phung, and Svetha Venkatesh. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of Biomedical Informatics*, 54:96–105, 2015.
- [31] Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S Jacobson. Noble-flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):32, 2016.
- [32] U.S. National Library of Medicine. Icd-9-cm diagnostic codes to snomed ct map, 2015.
- [33] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

- [34] Meng Wang. Predicting rich drug-drug interactions via biomedical knowledge graphs and text jointly embedding, 2017, arXiv preprint arXiv:1712.08875.
- [35] Meng Wang, Jiaheng Zhang, Jun Liu, Wei Hu, Sen Wang, Xue Li, and Wenqiang Liu. Pdd graph: Bridging electronic medical records and biomedical knowledge graphs via entity linking. In *International Semantic Web Conference*, pages 219–227. Springer, 2017.
- [36] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20, 2018.
- [37] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.
- [38] Jun Xu, Yaoyun Zhang, Hua Xu, et al. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176, 2015.
- [39] Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. A neural multi-task learning framework to jointly model medical named entity recognition and normalization, 2018, arXiv preprint arXiv:1812.06081.