Neural Networks 121 (2020) 132–139

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks

Luka Gligic, Andrey Kormilitzin *, Paul Goldberg, Alejo Nevado-Holgado

University of Oxford, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 20 February 2019
Received in revised form 29 July 2019
Accepted 29 August 2019
Available online 6 September 2019

Keywords:
Neural Networks
NLP
Named entity recognition
Electronic health records
Transfer learning
LSTM

ABSTRACT

Neural networks (NNs) have become the state of the art in many machine learning applications, such as image, sound (LeCun et al., 2015) and natural language processing (Young et al., 2017; Linggard et al., 2012). However, the success of NNs remains dependent on the availability of large labelled datasets, such as in the case of electronic health records (EHRs). With scarce data, NNs are unlikely to be able to extract this hidden information with practical accuracy. In this study, we develop an approach that solves these problems for named entity recognition, obtaining 94.6 F1 score in I2B2 2009 Medical Extraction Challenge (Uzuner et al., 2010), 4.3 above the architecture that won the competition. To achieve this, we bootstrap our NN models through transfer learning by pretraining word embeddings on a secondary task performed on a large pool of unannotated EHRs and using the output embeddings as a foundation of a range of NN architectures. Beyond the official I2B2 challenge, we further achieve 82.4 F1 on extracting relationships between medical terms using attention-based seq2seq models bootstrapped in the same manner.

Crown Copyright © 2019 Published by Elsevier Ltd. All rights reserved.

1. Introduction

Electronic Health Records (EHRs) are the databases used by hospital and general practitioners to daily log all the information they record from patients (Johnson, Fraser, Wyatt, & Walley, 2014). This information typically includes, but is not limited to: disorders, taken medications, dosages, symptoms, results from medical tests, and even considerations made by the doctor when evaluating each patient. In number of subjects (for example, 50 million patients in the case of European Medical Information Framework (EMIF)), EHRs are the largest source of empirical data in biomedical research (Denis, 2017; Jensen, Jensen, & Brunak, 2012), making them ideal for studying disease (e.g. Alzheimer's Perera, Khondoker, Broadbent, Breen, & Stewart, 2014, cardiovascular disease Perera et al., 2017, or associated risk factors Savova, Ogren, Duffy, Buntrock, & Chute, 2008; Stubbs & Uzuner, 2015; Uzuner, Goldstein, Luo, & Kohane, 2008) and evaluating service (e.g. monitoring adverse drug reactions (Iqbal et al., 2015)). However, most of the information held in EHRs is in the form of natural language text (i.e. written by the physician during each session with each patient), making it inaccessible for research (Jensen et al., 2012; Murdoch & Detsky, 2013). Unlocking all this information would represent a considerable contribution to biomedical research, multiplying the quantity and variety of scientifically useable data, which is the reason why major efforts have been relatively recently initiated towards this goal (Denis, 2017; Jackson M.Sc et al., 2014; Jensen et al., 2012; Savova et al., 2008) as well as being the main motivation behind this work.

The central idea of the paper is to develop an accurate and robust neural model for information extraction from medical texts, specifically, we were interested in medical named entity recognition (NER) and relation extraction (RE) between them. Although traditional Natural Language Processing (NLP) algorithms, such as rule systems (Karystianis et al., 2017), can perform this task with fair accuracy in the simpler situations (well-structured text, large amounts of labelled data available and many annotated samples), the challenge remains an unsolved problem in the more complex cases (badly structured language, few labelled samples) (Cambria & White, 2014). Unfortunately, data found in EHRs falls under the second category. Namely, physicians tend to use badly formatted shorthand and non-widespread acronyms ('transport pt to OT tid via W/C' for 'transport patient to occupational therapy three times a day via wheel chair'), while labelled records are scarce (ranging in the hundreds for a given task and with very few annotated samples). A reason for this scarcity is that data access is difficult due to ethical concerns (Entzeridou, Markopoulou, & Mollaki, 2018; Jamshed, Ozair, Sharma, & Aggarwal, 2015; Layman, 2008). Other reason is that, even with data access granted, medical text needs to be annotated by field experts (e.g. clinicians), who are themselves in short supply.

* Corresponding author.
E-mail addresses: luka.gligic@gtc.ox.ac.uk (L. Gligic), andrey.kormilitzin@psych.ox.ac.uk (A. Kormilitzin), paul.goldberg@cs.ox.ac.uk (P. Goldberg), alejo.nevado-holgado@psych.ox.ac.uk (A. Nevado-Holgado).

https://doi.org/10.1016/j.neunet.2019.08.032
0893-6080/Crown Copyright © 2019 Published by Elsevier Ltd. All rights reserved.

Table 1
I2B2 datasets used in this study. Third column indicates the total number of documents in each corpus. Fourth and fifth columns indicate which not annotated and annotated documents, respectively, were unique, and therefore added into the common pool of documents used for subsequent analyses and unsupervised and supervised training.

Year	Existing annotations	Total documents	Unique documents (not annotated)	Unique documents (annotated)
2007	Smoking	2886	926	0
2008	Obesity	1267	1237	0
2009	Medications	1945	991	258
2010	Term relations	696	694	0
2011	Conference	424	188	0
2012	Temporal relations	671	311	0
	Total	7889	**4347**	**258**

In the study presented in this paper we address these problems by: first, using Neural Networks (NN) (LeCun, Bengio, & Hinton, 2015; Linggard, Myers, & Nightingale, 2012), which are expected to be more robust to badly structured language than rules or other traditional techniques (Young, Hazarika, Poria, & Cambria, 2017); second, rather than training them only on the objective task, we bootstrap the Neural Networks through transfer learning, by feeding them pretrained word embeddings from a secondary task on unannotated electronic records. This approach achieves 94.7 F1 in I2B2 2009 Medical Information Extraction challenge, 4.3 more than the traditional approach that originally won the challenge. In addition to the official objectives of I2B2 2009, this approach also obtained 82.4 F1 on extracting the relationships between medical terms, which are of high importance in research with EHRs.

2. Methods

2.1. Objective task

Our objective task consisted on automatically locating and predicting the annotations of I2B2 2009 Medical Information Extraction challenge (Uzuner, Solti, & Cadag, 2010). These labels consisted of all mentions of medications where the patient was the user, plus a number of associated fields per term. These fields were: medication, dosage, mode, frequency, duration, reason. Medication includes compound name, brand name, generics, collectives and prescriptions (e.g. acetylsalicylic acid or aspirin). Dosage indicates the amount administered to the patient, which could be a measurement (e.g. 2.0 mgs) or units (e.g. 2 tablets). Mode refers to the administration route (e.g. orally). Frequency refers to how often the medication was taken (e.g. 2 per day). Duration consists on treatment length (e.g. until symptoms disappear). Reason is the cause for the prescription (e.g. presumed pneumonia).

2.2. Datasets

This study used all datasets released by I2B2 from 2007 to 2012. We observed that some documents were repeated across different yearly releases. To eliminate duplicates, we sequentially pooled each corpus into a final set of 4605 unique documents (see Table 1). I2B2 2009 challenge released a total of 1249 unique documents, with 258 of them annotated for the objective task. Given that our objective task was the one corresponding to I2B2 2009 challenge, only the 258 documents from this year were considered annotated for our case, using all others as unannotated samples for the purpose of transfer learning. In detail, 4347 unannotated samples were selected for training embeddings, 238 for training the rest of the NN, 10 for validation and 10 for final testing.

2.3. Text pre/processing

Text was pre-processed to reduce the number of out of vocabulary (OOV) words, which was defined as words not accounted by the embeddings described in Section 2.4. Sentences were split on "." followed by a capital letter, as recommended by Patrick and Li (2009). All numbers were replaced by the special token <num> . Punctuation symbols ".", ":" and ";" were removed, unless they were surrounded by letters or followed a number. All letters were lower cased. Pre-processing did not alter number and location of words and sentences. Finally, a number of metrics evaluated the text demographics of the embedding/train/validation/test datasets after pre-processing.

2.4. Training embeddings

We created two embeddings versions with Contiguous Bag of Words (CBOW) and Continuous Skip-Gram (CSG) (Mikolov, Chen, Corrado, & Dean, 2013a; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013b), and evaluated their adequacy for the objective task described in Section 2.1. Following the CBOW algorithm, we randomly initialised m-dimensional embeddings with a Gaussian distribution of mean 0 and standard deviation 1. The text of all samples (including not annotated and annotated, but excluding the 20 samples reserved for validation and final testing; see Section 2.2) was then randomly divided into 4.5 million windows of 11 words length each. Each window would contain only words from the same sentence of the central word, using a neutral 'PAD' symbol for positions that spread to other neighbouring sentences. A fully connected single layered network was then created to predict the central word of each window based on the average of all word embeddings appearing within the window. Using this network, embeddings were trained through backpropagation with 0.025 (min alpha 0.0001) learning rate, 5 epochs, and all other parameters set to default values of Word2Vec implementation from the *gensim* library (Řehůřek & Sojka, 2010). Separately to CBOW, and following the CSG algorithm, we initialised other set of 100 dimensional embeddings with a Gaussian distribution of mean 0 and standard deviation 1. Text from all not annotated samples was divided into windows in the same manner as done for CBOW. A fully connected single layer network was then trained through with 0.025 (min alpha 0.0001) learning rate and 5 epochs to predict words from the window based on the central embedding. In both cases, the size of the vocabulary consisted on all the words from the embedding and training sets (see Fig. 2).

2.5. Intrinsic evaluation of embeddings

Once created, we intrinsically (Bakarov, 2018) evaluated the embeddings by calculating their average Euclidean distance, average cosine similarity, and visualising their t-SNE projection. For the first of these, we divided all words into those belonging to each of the target categories (i.e. medication, dosage, mode, frequency, duration, reason; see Section 2.1), and those belonging to none. Then we calculated the average Euclidean distance between words of the same class. We followed the same process to calculate the average cosine similarity, but using cosine distance rather than Euclidean distance. Finally, word categories were projected onto a two-dimensional space with t-SNE and then visually inspected to assess class separation (Hinton & Bengio, 2008).

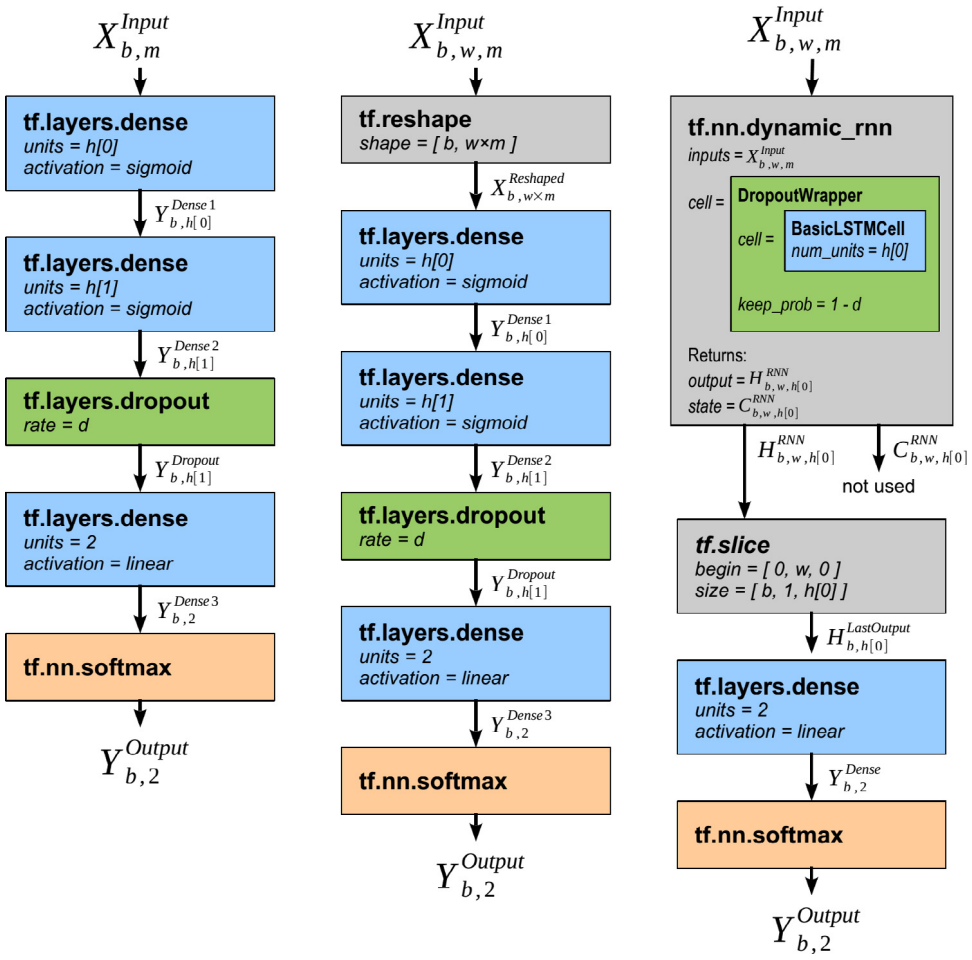

Fig. 1. Architectures for term classification. From left to right, the figure shows the context free FNN, the context aware FNN and the RNN architectures used for terms classification. The component operations (e.g. layers) of each architecture are represented as boxes, with blue for full layers, green for dropout, orange for transformation functions, and grey for shuffling or tensorflow wrappers. Within each box, bold font shows the name of the tensorflow operation, and italic fonts the input parameters when non default values were used for that particular operation. In occasions, input and output tensors are also represented with a capital letter, with subindex for tensor dimensions, and superindex for a further description of the data held in that particular tensor.

2.6. Extrinsic evaluation of embeddings

Besides the three intrinsic evaluation methods described in Section 2.5, we also extrinsically evaluated them with a context free classification task (Bakarov, 2018). The task consisted on classifying words as either belonging to each of the target classes of the study (i.e. medication, dosage, mode, frequency, duration, reason; see Section 2.1) or to none. The task was implemented in the form of a series of binary classifiers, one independently for each target class, and results averaged. The classifier was a feed forwards neural network (FFN) whose input was only the m-dimensional embedding of the to-be-classified word, followed by 'l' densely connected sigmoid layers of 'h' units each, and finally a dense SoftMax layer of 2 units, corresponding with the one-hot representation of the classification objective. Each of the 'h' dense layers was also followed by a dropout operation with proportion 'd' per cent. In the context of this article, we will call this architecture "context free FFN". The training and testing sets were 10 000 and 1000 randomly selected words, with 'p'% of them belonging to one of the target classes of the study. The NN was trained with Adam for 'e' epochs, learning rate 'r', using batches of size 'b'. Several values of parameters 'm', 'l', 'h', 'd', 'p', 'e', 'r' and 'b' were tested to prevent using an architecture, dataset or training method that specially favoured either CBOW or CSG.

2.7. Term classification

The "context free FFN"[1] defined in Section 2.6 was also used to obtain a baseline measure of performance on the objective task (Section 2.1) with the objective dataset (2.2). In this case we set all free parameters ('m', 'l', 'h', 'd', 'p', 'e', 'r' and 'b') to the values that produced the best performance on the set of words randomly selected in Section 2.6.

A second architecture was created by extending the context free FFN into a "context aware FFN".[2] This consisted on replacing the single word input by the concatenation of the 'w' words existing around the to-be-classified token. Namely, the one-dimensional embedding, which consisted of 'm' real numbers each, was concatenated into a single 1D vector of 'm(1+2w)' real numbers.

A third architecture, partly based on previous work (Mesnil et al., 2015), was a "RNN"[3] (recurrent neural network) that sequentially read all words in the target window around the target word. The input to the architecture was one word embedding per time step, fully connected to a LSTM layer of 100 units. The final

[1] In the GitHub repository, this architecture is defined in file 'Model 2 (Feed Forward).ipynb'.

[2] Defined in file 'Model 3 (Windowed Feed Forward).ipynb'.

[3] Defined in 'Model 4 (Recurrent).ipynb'.

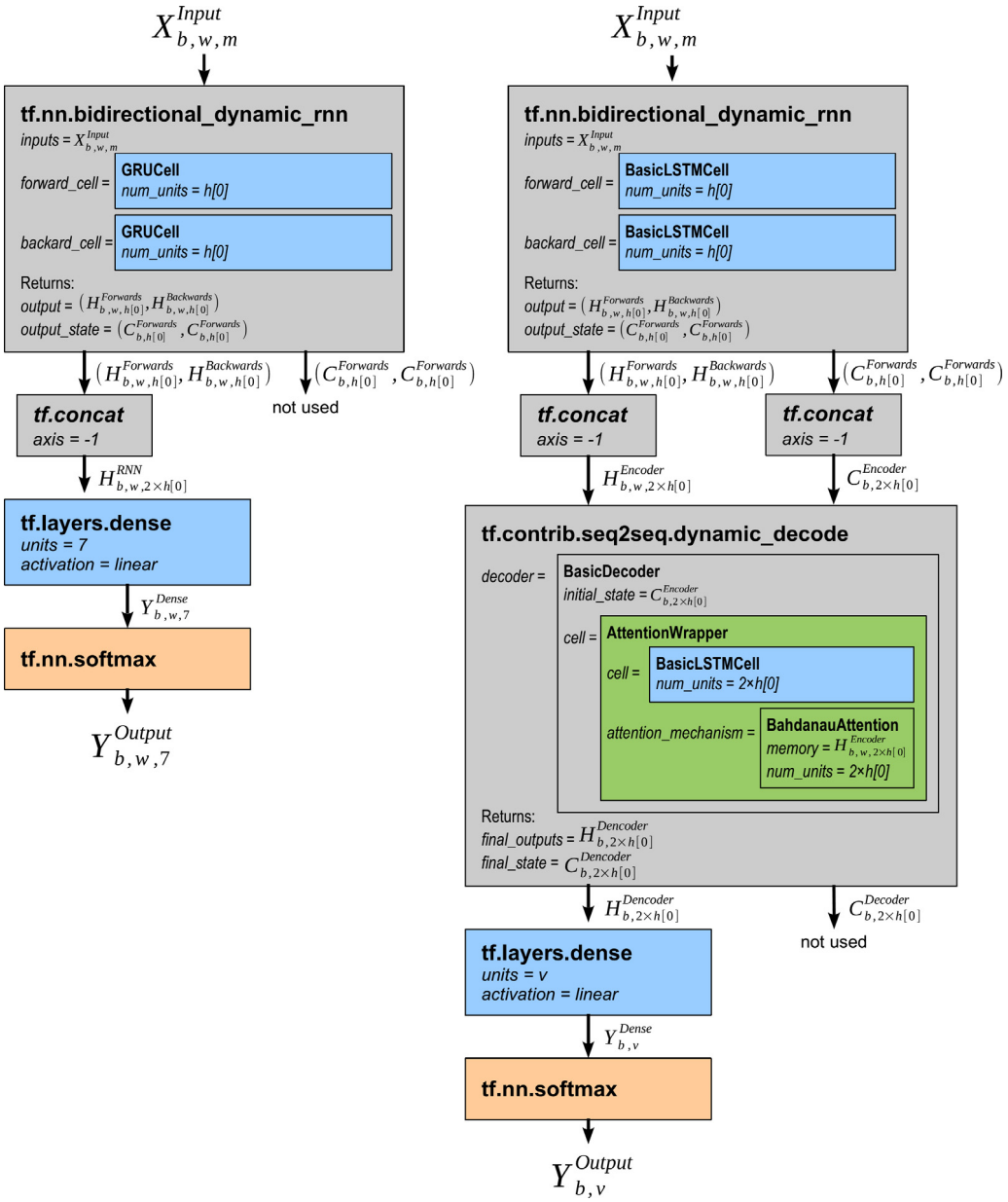

Fig. 2. Architectures for relationship extraction. From left to right, the figure shows the seq2seq and the encoder–decoder RNN architectures. Boxes, colours and fonts have same meaning as in Fig. 1.

state of the LSTM layer is fed to a SoftMax function. The NN was trained via Adam algorithm, 0.001 learning rate, 50 batch size, 3 epochs.

2.8. Relationship extraction

I2B2 challenge consisted on extracting all medications, dosages, modes, frequencies, durations and reasons as individual terms (see Section 2.1), and the architectures of Section 2.7 were designed and tested for this objective. However, in practice, what is of importance is not only the medical terms themselves, but also the relationships between them. Namely, when extracted medical information is used in a subsequent epidemiological analysis, it is of little value to know that a patient took, for example, aspirin, as this patient could have taken the drug on only one occasion, which would have no long-term impact on chronic diseases. What in that example would be of interest is to know whether the patient takes aspirin daily, for how long and with what dosage. Therefore, due to the importance of extracting relationships between medical terms, we also designed and tested a fourth and a fifth architecture specialised on, given a target medication term, extracting its dosage, mode, frequency, duration and reason.

The fourth architecture, which was the first one used for this task, was a sequence to sequence (seq2seq) RNN,[4] which sequentially read all words within a 5 rows window around the target medication word, simultaneously outputting word classification at each time step. A bidirectional neural network architecture comprising 100 gated recurrent units (GRU) was initialised with a linear transformation of bag of words (BOW) representation of the target medication for that window. This BOW representation consisted on the sum of all words part of the target medication term (e.g. for the term 'baby aspirin', embedding of 'baby' plus embedding of 'aspirin') concatenated with the medication label,

[4] Defined in 'Model 11 (ELS2S).ipynb'.

Table 2
Document metrics of annotated datasets. The table shows how many documents/entries/phrases/tokens correspond to each of the 3 annotated datasets (train/validation/test) used.

Metric	Train	Validation	Test
No. of documents	238	10	10
No. of entries	8387	485	376
No. of phrases	21497	1329	973
No. of tokens	34718	2169	1571
Mean entries per document	35.2	48.5	37.6
Mean phrases per document	90.3	132.9	97.3
Mean tokens per document	145.9	216.9	157.1
Mean phrases per entry	2.6	2.7	2.6
Mean tokens per entry	4.1	4.5	4.2
Mean tokens per phrase	1.6	1.6	1.6
Vocabulary of target tokens	2267	442	4372
Out of vocabulary tokens	N/A	48	52

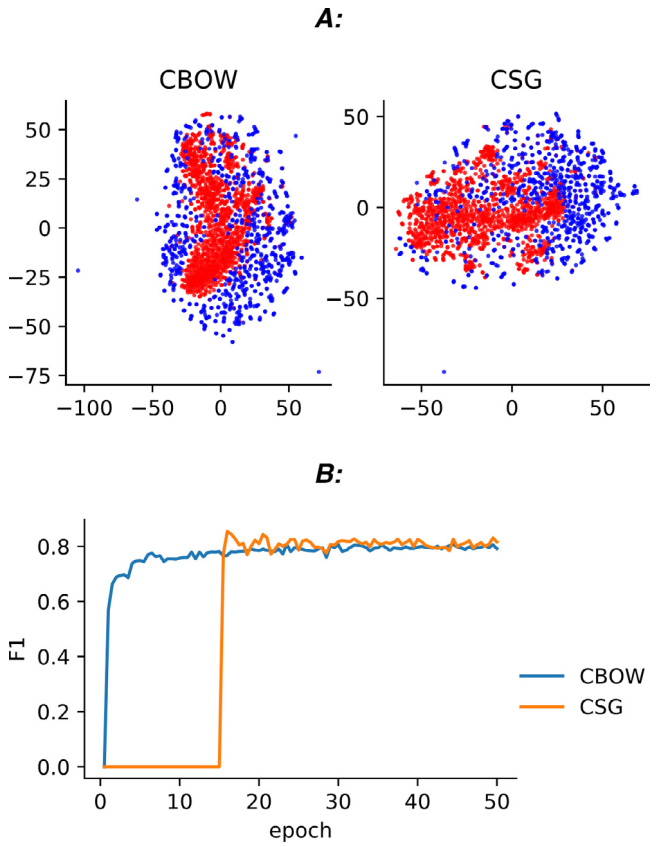

Fig. 3. Evaluation of embeddings. **A**: Intrinsic evaluation with t-SNE. The figure shows the 2D t-SNE projection of the embeddings calculated with either CBOW (left) or CSG (right). Each point is an embedding, with red corresponding to target categories and blue to other words. **B:** Extrinsic evaluation. The figure shows the F1 score of the context free NN when embeddings are trained using CBOW (blue) or CSG (orange) algorithms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which altogether created a vector of length 'm' (size of embeddings) plus 1 (for the medication label). The weights and biases of the linear transformation were learnt during training. Then, the GRU was sequentially fed with the 100-dimensional word embeddings of the sentence, where embeddings were concatenated with an additional real number representing the I2B2 2009 classification of each word, if any (i.e. 1 for medication, 2 dosage, 3 mode, 4 frequency, 5 duration, 6 reason and 0 for 'none'). In each time step, the state of the GRU was fed to a SoftMax layer of 7 outputs, representing each of the I2B2 2009 term classes (plus a 7th class for 'none'). The RNN was trained via Adam algorithm, 0.001 learning rate, 50 batch size, 100 epochs (see Table 5).

The fifth and final architecture, which was the second one used for the relationships task, was an encoder–decoder RNN,[5] which first read all words within a ± 2 row window and then outputted all those words deemed as related to the target medication. A bidirectional LSTM encoder of 128 units was initialised with a BOW representation of the target medication. Then, in the encoding phase, the LSTM read the input window coded as in the seq2seq RNN model described in the previous paragraph. On reaching the end of the window, the final states of the encoder in the forwards and backwards directions are concatenated to form the initial 256-dimensional state of a decoder LSTM. During the decoding phase, this second LSTM received as input the step outputs of the decoder weighted by either Bahdanau, Cho, and Bengio (2014) or Luong, Pham, and Manning (2015) attention mechanism. The decoding LSTM then outputted words until emitting a special <end of output> token. Output words were selected with a SoftMax over the whole vocabulary. The RNN was trained via Adam algorithm with power scheduling rate decay, 0.001 learning rate, 0.00001 decay rate, gradients clipped at value 5, 50 batch size, 100 epochs.

In the case of the latter architecture (encoder–decoder RNN), it should be noted that as the model itself produces words rather than labels, it is impossible to assess its results for field specific Type I errors, so a vocabulary lookup function was used to determine the fields of false positive tokens.

3. Results

3.1. Text pre-processing

Each document contains a number of entries, which are further divided into sentences and tokens. A number of document metrics count how documents/entries/sentences/tokens correspond to each other. The total number of unique tokens appearing in the unannotated dataset (see Table 2) forms the vocabulary size of our embeddings, which does not include a small number of words of the validation (5) and testing (7) sets. Further labels metrics indicate that pre-annotated terms are evenly distributed across train, validation and testing sets (see Table 2).

3.2. Intrinsic evaluation of embeddings

Intrinsic evaluation did not clearly favour one method of constructing embeddings above the other (see Table 4). Average Euclidean distance showed preference for CSG embeddings over CBOW, while average cosine similarity did the opposite. Visual inspection with t-SNE (see Fig. 3) indicated that both methods separated words belonging to target categories (i.e. medication, dosage, mode, frequency, duration, reason; see Section 2.1) from the rest, but again without a method clearly outperforming the other. We also explored with embedding sizes of 2^4 to 2^{10} and noticed diminishing improvements in performance at values above 2^7, ultimately settling at an embedding size of 100.

3.3. Extrinsic evaluation of embeddings

To further evaluate embeddings, we created a context free FFN whose input was the embedding of a single word and trained it on classifying such words as either belonging to any of the target classes of the study or to none (see Section 2.6). The NN meta-parameters that we explored and the values that obtained best performance are in Table 6. One single layer, sigmoid activation

[5] Defined in 'Model 10 (S2S).ipynb'.

Table 3
Label metrics of our annotated datasets. The table shows the proportions of entries that contain each of the I2B2 2009 labels.

Field	Train	Validation	Test
Medication	100%	100%	100%
Dosage	49.5%	56.3%	50.0%
Mode	37.7%	40.8%	37.7%
Frequency	44.8%	53.4%	45.4%
Duration	6.1%	6.0%	6.1%
Reason	18.3%	17.5%	18.1%

Table 4
Intrinsic evaluation of embeddings. The table shows results of the intrinsic evaluation performed on the embeddings trained either with CBOW or CSG (see Section 2.5). AED — Average Euclidean Distance; ACD — Average Cosine Similarity.

Field	CBOW		CSG	
	AED	ACS	AED	ACS
Medication	6.53	0.23	3.61	0.53
Dosage	11.93	0.15	4.45	0.42
Mode	10.26	0.21	4.51	0.43
Frequency	14.63	0.15	4.76	0.41
Duration	17.01	0.07	4.91	0.34
Reason	12.65	0.10	4.68	0.35

Table 5
Used parameters. The table shows the values used for the parameters of each architecture.

Parameter	Context free FFN	Context aware FFN	RNN
m = embeddings dimension	100	100	100
w = window words	–	5	15
l = No. of layers	2	2	1
h = No. of units per layer	[100, 100]	[500, 100]	(100)
d = dropout proportion	0.0	0.0	0.0
p = proportion of target words	0.1	0.1	0.1
e = No. of epochs	5	5	3
r = learning rate	0.01	0.001	0.001
d = decay rate	0.002	0.0	0.0
b = batch size	50	50	50

Table 6
Metaparameters of context free NN. The table shows the best performing set of parameters for each field.

Parameter	Explored values	Med	Dos	Mod	Fre	Dur	Rea
Algorithm	CBOW, CSG	CBOW	CSG	CSG	CBOW	CSG	CBOW
No. of layers 'l'	1, 2	1	1	1	1	1	1
Activation 'a'	tanh, σ, ReLU	σ	σ	σ	σ	σ	σ
Dropout 'd'	0.0, 0.2, 0.4	0.0	0.4	0.2	0.4	0.4	0.4
Lean rate 'r'	0.001, 0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 7
Performance on I2B2 2009 objective task. The table shows F1 scores for each of our three architectures on extracting each of the target terms of I2B2 2009. For comparison, results of the winners of I2B2 challenge are also provided in the last column.

Term	Context free FFN	Context aware FFN	RNN	I2B2 winner
Medication	79.0	88.9	**94.6**	90.3
Dosage	71.0	91.0	**93.0**	90.8
Mode	95.4	92.7	**96.9**	89.3
Frequency	79.8	88.5	**90.9**	87.7
Duration	31.7	61.9	**63.0**	56.0
Reason	26.5	28.1	28.4	**47.0**

functions, dropout at 0.4 and a learning rate of 0.01 obtained in general the best performance. However, no significant difference in F1 was found between CBOW and CSG algorithms (see Fig. 3), although CBOW converged earlier and had a more stable final performance (see Fig. 4).

Table 8
Performance on relationship extraction task. The table shows F1 scores for each of our architectures capable of extracting relationships between I2B2 2009 terms and pre-annotated drugs.

Term	seq2seq RNN	Encoder–decoder RNN + Bahdanau	Encoder–decoder RNN + Luong
Average	0.824	0.806	0.811
Medication	0.897	0.851	0.876
Dosage	0.797	0.876	0.879
Mode	0.863	0.889	0.831
Frequency	0.811	0.785	0.826
Duration	0.701	0.434	0.547
Reason	0.667	0.463	0.402

3.4. Term classification

Three architectures were trained and tested on the objective task of the original I2B2 2009 challenge. The first architecture was the context free FFN described in Section 2.6 with the optimal metaparameter values of Section 3.3. The second architecture was a context aware FFN, which extended the previous context free architecture by also reading the '±w' words existing around the to-be-classified token. The third architecture was a LSTM-based RNN capped by a SoftMax that sequentially read the '±w' words existing around the target token. This last architecture outperformed the FNN models in all target terms. Further its performance was above the winner algorithm of I2B2 2009 challenge in all tasks except for extracting 'reason' (see Table 7). Interestingly, context aware FFN preferred small window sizes, while the performance of the RNN was not specially affected by the value of 'w' (see supplementary Figure 4).

3.5. Relationship extraction

Beyond the official I2B2 2009 term extraction task, we also created two architectures to identify all terms associated to a given pre-annotated drug (see Section 2.8). These were a seq2seq RNN, which simultaneously read the input word by word while outputting word classification, and an encoder–decoder RNN, which first read all input words and then outputted all those related to the pre-annotated drug. The encoder–decoder system was trained and tested with two different methods of attention — Bahdanau et al. (2014) and Luong et al. (2015). Examples of extractions by these architectures are shown in Fig. 5 and the results can be seen in Table 8.

4. Discussion

Architectures based on the artificial neural networks suffer from requiring large amounts of annotated data to be able to perform at state-of-the-art-accuracy. This fact bars them from applications where data is scarce or difficult to access and annotate, such as EHRs. This is the reason why laboratories working with EHRs have traditionally preferred classical methods such as rule-based systems (Cunningham, Tablan, Roberts, & Bontcheva, 2013; Karystianis et al., 2017; Perera et al., 2014). In this study we demonstrate that appropriate use of transfer learning and unsupervised learning allow NNs to perform above traditional methods such as those applied earlier (Uzuner et al., 2010). Specifically, fine tuning embeddings to domain specific text (i.e. medical text) and the use of recurrent architectures appeared to produce the highest gains in performance. Interestingly, high dropout rates performed better than low dropout rates only for the terms that were least annotated (see Table 3), even when the most densely annotated terms (e.g. 'medication') were only sampled in 238 documents.

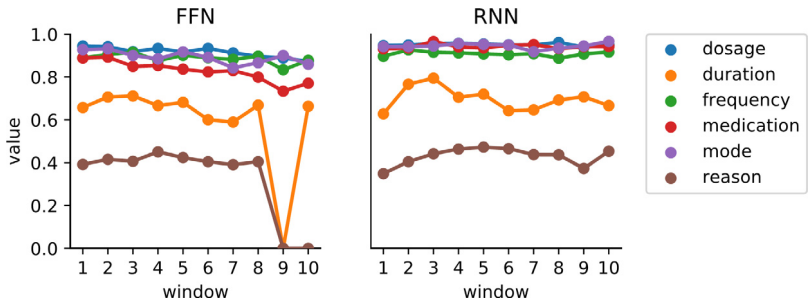

Fig. 4. Effects of window size. The figure shows how F1 varies depending on window size of the context aware FFN (left) and the RNN (right) architectures.

Word embedding input	vancomycin <start> her graft the remainder of the hospital course was unremarkable on, the <num> of july , she was discharged back to the hospital discharge medication vancomycin <num> mg iv q d , ofloxacin <num> mg po bid (both antiotics to continue for an additional to week course) , Coumadin with target
Word class input	1 0 6 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 2 3 4 4 0 1 2 2 3 4 0 0 1 0 0 5 5 5 5 5 5 0 0 1 0 0
Output	Vancomycin <num> mg iv q d for an additional two week course <eos>

Fig. 5. Term relationship sample. The table shows the two streams of input (word embedding and word class) that both the seq2seq and the encoder–decoder RNNs would receive in this sample. The third row shows the output given by the encoder–decoder after it read this particular example, while the last row shows the ground truth.

However, our model still did perform poorly for the least annotated categories (e.g. 'reason', see Table 3), where the traditional knowledge-based approaches that won the original challenge achieved better results (see Table 7). The same problem arose for relationship extraction (Section 3.5), because each sample was now each record entry (e.g. each record with a word of the category 'medication'), rather than each annotated word (e.g. each word of the category 'medication'), as implied in Fig. 5.

Future work could attempt at further improving the performance of NNs in small annotated datasets by transferring learning from unannotated datasets larger than what we used here, and using both within-domain (e.g. medical) and out-of-domain corpora. It is striking that a non-medical expert can learn to recognise reasons for prescribing medications (i.e. our category 'reason') in EHRs after only seeing a few examples, while NNs still reach only F1 score of 0.281 even after seeing numerous more examples than a human. To mitigate the problem of learning from scarce data, a few-shot learning approach for medical texts was introduced recently (Hofer, Kormilitzin, Goldberg, & Nevado-Holgado, 2018). One of the challenges outlined in above, namely the representation of the worst performing categories, such as 'reasons', could be addressed using fuzzy sets and fuzzy logic due to their ability to capture semantics of vague linguistic constructs due to their capacity (Zadeh, 1996). This approach was studied in recent works (Qiu et al., 2019; Sun et al., 2018), where an adaptive fuzzy control scheme for stochastic non-linear systems was introduced as well as using an efficient representation of high-dimensional ordered data using the path signature from stochastic analysis (Chevyrev & Kormilitzin, 2016; Kormilitzin et al., 2016, 2017; Lyons, 2014). Given that knowledge-based methods still outperformed our NN in the 'reason' category (F1 = 0.47), other avenues could consist on introducing field knowledge into the NN in the form of bias, or in the form of symbolic methods such as dictionaries and gazetteers. Finally, more theoretical work such as the Information Bottleneck (Tishby & Zaslavsky, 2015), the Neural Homology (Guss & Salakhutdinov, 2018), or other theories could allow us to better understand why NNs still need such a large number of samples to learn appropriately, and guide future work on how this problem could be overcome.

Acknowledgements

The study was funded by the National Institute for Health Research's (NIHR), United Kingdom Oxford Health Biomedical Research Centre (BRC-1215-20005). This work was supported by the UK Clinical Records Interactive Search (UK-CRIS) system funded and developed by the NIHR Oxford Health BRC at Oxford Health NHS Foundation Trust and the Department of Psychiatry, University of Oxford. The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), the NIHR, the MRC or the Department of Health. AK, ANH are funded by the Medical Research Council (MRC), United Kingdom, Pathfinder Grant (MC-PC-17215).

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, ArXiv14090473 Cs Stat. http://arxiv.org/abs/1409.0473 (accessed 05.11.18).

Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods, ArXiv180109536 Cs. http://arxiv.org/abs/1801.09536 (accessed 05.11.18).

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Computer Intelligence and Magazine*, *9*, 48–57. http://dx.doi.org/10.1109/MCI.2014.2307227.

Chevyrev, Ilya, & Kormilitzin, Andrey (2016). A primer on the signature method in machine learning. arXiv preprint arXiv:1603.03788.

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, *9*, e1002854. http://dx.doi.org/10.1371/journal.pcbi.1002854.

Denis, M. (2017). U.K. Clinical record interactive search (CRIS). *Alzheimers Dementia*, *13*, P1223. http://dx.doi.org/10.1016/j.jalz.2017.07.413.

Entzeridou, E., Markopoulou, E., & Mollaki, V. (2018). Public and physician's expectations and ethical concerns about electronic health record: Benefits outweigh risks except for information security. *International Journal of Medical Information*, *110*, 98–107. http://dx.doi.org/10.1016/j.ijmedinf.2017.12.004.

Guss, W. H., & Salakhutdinov, R. (2018). On Characterizing the Capacity of Neural Networks using Algebraic Topology, ArXiv180204443 Cs Math Stat. http://arxiv.org/abs/1802.04443 (accessed 26.11.18).

Hinton, G., & Bengio, Y. (2008). Visualizing data using t-SNE. In *Cost-sensitive mach. learn. inf. retr. 33, n.d.* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.457.7213.

Hofer, M., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2018). Few-shot Learning for Named Entity Recognition in Medical Text. arXiv preprint arXiv:1811.05468.

Iqbal, E., Mallah, R., Jackson, R. G., Ball, M., Ibrahim, Z. M., Broadbent, M., et al. (2015). Identification of adverse drug events from free text electronic patient records and information in a large mental health Case register. *PLoS One*, *10*, e0134208. http://dx.doi.org/10.1371/journal.pone.0134208.

Jackson M.Sc, R. G., Ball, M., Patel, R., Hayes, R. D., Dobson, R. J., & Stewart, R. (2014). Texthunter – a user friendly tool for extracting generic concepts from free text in clinical research. In *AMIA. Annu. Symp. Proc., Vol. 2014* (pp. 729–738).

Jamshed, N., Ozair, F., Sharma, A., & Aggarwal, P. (2015). Ethical issues in electronic health records: A general overview. *Perspectives Clinical Research*, *6*, 73. http://dx.doi.org/10.4103/2229-3485.153997.

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Natural Review Genetics*, *13*, 395–405. http://dx.doi.org/10.1038/nrg3208.

Johnson, O. A., Fraser, H. S. F., Wyatt, J. C., & Walley, J. D. (2014). Electronic health records in the UK and USA. *The Lancet*, *384*, 954. http://dx.doi.org/10.1016/S0140-6736(14)61626-3.

Karystianis, G., Nevado, A. J., Kim, C.-H., Dehghan, A., Keane, J. A., & Nenadic, G. (2017). Automatic mining of symptom severity from psychiatric evaluation notes. *International Journal of Methods and Psychiatry Research*, e1602. http://dx.doi.org/10.1002/mpr.1602.

Kormilitzin, A. B., et al. (2016). Application of the signature method to pattern recognition in the cequel clinical trial. arXiv preprint arXiv:1606.02074.

Kormilitzin, Andrey, et al. (2017). Detecting early signs of depressive and manic episodes in patients with bipolar disorder using the signature-based model. arXiv preprint arXiv:1708.01206.

Layman, E. J. (2008). Ethical issues and the electronic health record:. *Health Care Management*, *27*, 165–176. http://dx.doi.org/10.1097/01.HCM.0000285044.19666.a8.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. http://dx.doi.org/10.1038/nature14539.

Linggard, R., Myers, D. J., & Nightingale, C. (2012). *Neural networks for vision, speech and natural language*. Dordrecht: Springer Netherlands, http://public.eblib.com/choice/publicfullrecord.aspx?p=3565560 (accessed 21.07.17).

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation, ArXiv150804025 Cs. http://arxiv.org/abs/1508.04025 (accessed 05.11.18).

Lyons, Terry (2014). Rough paths, signatures and the modelling of functions on streams. arXiv preprint arXiv:1405.4537.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., et al. (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEEACM Transactions on Audio, Speech and Language Processing*, *23*, 530–539. http://dx.doi.org/10.1109/TASLP.2014.2383614.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, ArXiv13013781 Cs. http://arxiv.org/abs/1301.3781 (accessed 05.11.18).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality, ArXiv13104546 Cs Stat. http://arxiv.org/abs/1310.4546 (accessed 05.09.16).

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health Care. *JAMA*, *309*, 1351. http://dx.doi.org/10.1001/jama.2013.393.

Patrick, J., & Li, M. (2009). High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of American Medical and Information Association*, *17*(2010), 524–527. http://dx.doi.org/10.1136/jamia.2010.003939.

Perera, G., Gungabissoon, U., Alexander, M., Ansel, D., Avillach, P., Salles, T. D., et al. (2017). Levels of blood pressure, body mass index and total serum cholesterol at different time points prior to dementia diagnosis: a case control study of over 28 million electronic health records from the emif ehr data resource. *Alzheimers Dementia*, *13*, P1420–P1421. http://dx.doi.org/10.1016/j.jalz.2017.06.2211.

Perera, G., Khondoker, M., Broadbent, M., Breen, G., & Stewart, R. (2014). Factors associated with response to acetylcholinesterase inhibition in dementia: A cohort study from a secondary mental health Care Case register in London. *PLoS One*, *9*, e109484. http://dx.doi.org/10.1371/journal.pone.0109484.

Qiu, Jianbin, et al. (2019). Observer-based fuzzy adaptive event-triggered control for pure-feedback nonlinear systems with prescribed performance. *IEEE Transactions on Fuzzy Systems*.

Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D., & Chute, C. G. (2008). Mayo clinic NLP system for patient smoking status identification. *Journal of American Medical and Information Association*, *15*, 25–28. http://dx.doi.org/10.1197/jamia.M2437.

Stubbs, A., & Uzuner, Ö. (2015). Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics*, *58*, S78–S91. http://dx.doi.org/10.1016/j.jbi.2015.05.009.

Sun, Kangkang, et al. (2018). Adaptive fuzzy control for non-triangular structural stochastic switched nonlinear systems with full state constraints. *IEEE Transactions on Fuzzy Systems*.

Tishby, N., & Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle, ArXiv150302406 Cs. http://arxiv.org/abs/1503.02406 (accessed 26.11.18).

Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of American Medical and Information Association*, *15*, 14–24. http://dx.doi.org/10.1197/jamia.M2408.

Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of American Medical and Information Association*, *17*, 514–518. http://dx.doi.org/10.1136/jamia.2010.003947.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proc. LREC 2010 workshop new chall. NLP framew* (pp. 45–50). ELRA, Valletta, Malta.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing, ArXiv170802709 Cs. http://arxiv.org/abs/1708.02709 (accessed 05.11.18).

Zadeh, Lotfi A. (1996). Quantitative fuzzy semantics. Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh. 105-122.