Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches

Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches

Philip John Gorinski[1], Honghan Wu[2], Claire Grover[1], Richard Tobin[1], Conn Talbot[3], Heather Whalley[3], Cathie Sudlow[2], William Whiteley[3], Beatrice Alex[1,4]
[1]Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh; [2]Usher Institute, University of Edinburgh; [3]Centre for Clinical Brain Sciences, University of Edinburgh; [4]Edinburgh Futures Institute, University of Edinburgh

Abstract

This work investigates multiple approaches to Named Entity Recognition (NER) for text in Electronic Health Record (EHR) data. In particular, we look into the application of (i) rule-based, (ii) deep learning and (iii) transfer learning systems for the task of NER on brain imaging reports with a focus on records from patients with stroke. We explore the strengths and weaknesses of each approach, develop rules and train on a common dataset, and evaluate each system's performance on common test sets of Scottish radiology reports from two sources (brain imaging reports in ESS – Edinburgh Stroke Study data collected by NHS Lothian as well as radiology reports created in NHS Tayside). Our comparison shows that a hand-crafted system is the most accurate way to automatically label EHR, but machine learning approaches can provide a feasible alternative where resources for a manual system are not readily available.

Introduction

Named Entity Recognition (NER) is an area of Natural Language Processing (NLP) that addresses the identification and classification of entities in written text. It has been employed using large variety of methods and on a multitude of domains and methods.[1–4]

Electronic Health Records (EHR) typically contain not only structured information about a patient but also written, unstructured text describing health professionals' opinions. Named entities in this domain include names of diseases, symptoms and anatomical locations. A radiology report is the opinion of a radiologist on a scan or X-ray. Figure 1 shows an example of an anonymised radiology report of a brain scan with identified named entities.

Report:
There is loss of the neuronal tissue in the left inferior frontal and superior temporal lobes, consistent with a prior infarct. There is generalised cerebral volume loss which appears within normal limits for the patient's age, with no focal element to the generalised atrophy. Major intracranial vessels appear patent. White matter of the brain appears largely normal, with no evidence of significant small vessel disease. No mass lesion, hydrocephalus or extra axial collection.
Summary:
Old left hemispheric infarct. No other significant finding.

location:cortical time:old ischaemic stroke atrophy small vessel disease tumour subdural haematoma

Figure 1: Example of a brain imaging report with annotated entities and their types below.

Related Work

NER is a well-studied field of NLP.[1,8,9] In 2003, Tjong and et al.[7] introduced a shared NER task at the Conference on Computational Natural Language Learning (CoNLL), which established a widely-accepted benchmark for the evaluation of NER systems. This led to research into machine learning methods, such as the Stanford NER tagger.[10] Other NER systems follow a rule-based approach, such as the ANNIE NER tagger.[11]

NLP for the medical domain has been an active field of research since the early 2000's. BioCreative and BioNLP provided shared tasks for NER and Relation Extraction (RE), with several systems applying NLP to biomedical text.[4,12–16] An overview of approaches to information extraction from EHR data was conducted by Meystre et al. (2008)[17], and Pons et al. (2016)[18] provide a recent review of NLP in radiology.

Most relevant for the systems investigated in the current work in terms of domain is work conducted by Flynn et al. (2010)[19], who present a system for the analysis of brain scan radiology reports. While not dealing with NER as a their main task, the authors applied keyword matching to analyse reports from the Tayside dataset, and assign document-level labels differentiating between stroke type (ischaemic stroke versus intracerebral haemorrhage).

There are many Machine Learning/Deep Learning architectures proposed in the literature for NER. In this paper, we draw from the line of work presented in Huang at al. (2015)[27], who employed Conditional Random Fields (CRF)[26] on top of bidirectional Long Short-Term Memories (LSTM).[5] Cornegruta et al. (2016)[6] evaluated a NER method on radiology reports. They employed a bidirectional LSTM (BiLSTM) neural network architecture, which they contrasted with a simple baseline of string matching against external lexicons.

Transfer learning[20] methods reuse machine learning models originally trained for a source task in a new target task. This idea has been adopted for doing NER tasks in a transferable manner, e.g. Arnold et al (2008)[22] used feature hierarchy, Nothman et al. (2013)[21] utilised the text and structure of Wikipedia, and Collobert and others created a convolutional neural network to jointly train multiple tasks.[23]

Data

The datasets we used to perform NER consist of anonymised radiology reports from brain MRI and CT scans conducted as part of the Edinburgh Stroke Study (ESS)[24] (n=1,168) and routine scans conducted by NHS Tayside (n=156,619). From the ESS data, a subset of 630 reports were annotated. From the Tayside collection, two subsets (Tay and TayExt) were selected and annotated. Each subset consists of reports for training/development of NER systems, as well as held-out test reports for testing and system comparison (see Table 1 for some statistics).

	ESS dev	ESS test	Tay dev	Tay test	TayExt dev	TayExt test
#Reports	364	266	362	700	1,068	300
#Sentences	3,731	2,755	2,791	3,948	8,401	2,677
#Named Entities	4,332	2,924	2,998	2,987	7,642	2,637

Table 1: Number of reports, sentences and named entities per subset and development/training (dev) and test splits.

ESS was the first set to be annotated by domain experts, and the rule-based EdIE-R system[25] was developed on this dataset. Data from NHS Tayside (Tay) was subsequently annotated with the same annotation scheme. This not only provides us with additional data, but also introduces different distributions of entities. This difference in data was further amplified by a second round of annotation on Tayside (TayExt), with reports specifically selected to include low-frequency entities.[1] Detailed frequency counts for entities annotated in ESS, Tay and TayExt are shown in Table 2.

Each set contains rich annotations of named entities in the text but also includes negated entities, entity relations and document-level labels. In this paper, we only focus on the entity annotation, not distinguishing between positive and negative entities. To ensure consistency, a first round of annotations from different annotators were compared before annotators carried out their work for the full datasets. Annotators showed very high inter-annotator agreement (IAA) on the test data sets (see *IAA* column in Tables 4 and 5 in the Experiments section). We report IAA figures for the entire ESS test data and for a subset of 100 reports from the Tayside test data.[2]

NER System Descriptions

For our comparative experiments on NER performance, we chose to evaluate a rule-based, a deep learning and a transfer learning system which we introduce here.

[1] TayExt is a subset of Tayside brain imaging reports which mention one of a list of keywords (e.g. bleed*, subarachnoid, subdural, haemorrh*, hemorrh*, mass, tumour or tumor). This filtering was done to ensure that certain entities which appeared infrequently in the previous datasets would be more frequent.

[2] We do not report results for the TayExt data because double annotation for that dataset has yet to be carried out.

Entity Type	ESS dev	ESS test	Tay dev	Tay test	TayExt dev	TayExt test
ischaemic stroke	697	455	369	306	668	214
haemorrhagic stroke	344	267	428	294	890	280
stroke	60	26	32	9	33	5
glioma tumour	0	0	10	9	32	12
meningioma tumour	4	8	9	2	32	6
metastasis tumour	24	12	61	120	117	35
tumour	297	166	146	303	432	117
subdural haematoma	244	109	75	95	968	309
small vessel disease	427	276	61	173	288	74
atrophy	246	153	105	168	350	90
microhaemorrhage	12	10	0	6	1	2
subarachnoid haemorrhage	13	10	50	16	135	54
haemorrhagic transformation	5	2	16	1	44	10
location:cortical	516	412	924	476	1775	665
location:deep	524	343	299	574	697	273
time:old	527	321	250	158	558	218
time:recent	392	354	163	277	622	273

Table 2: Per-entity frequency counts in the ESS, Tay and TayExt development (dev) and test datasets.

EdIE-R

EdIE-R (Edinburgh Information Extraction for Radiology reports)[25] is a rule-based system. It consists of a full pipeline that starts with the raw input text, and subsequently adds sectioning, tokenisation, sentence-splitting and linguistic annotation such as part-of-speech (POS) tagging and shallow syntactic analysis.[3] Figure 2 provides a schematic overview of the *EdIE-R* pipeline.

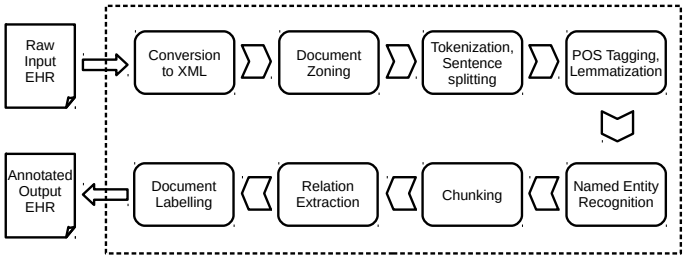

Figure 2: Overview of the *EdIE-R* pipeline.

Of particular interest to the comparative evaluation presented in this work is the NER step of *EdIE-R*. At this stage in the pipeline, the raw text has already been tokenised and POS tagged. Making use of hand-crafted rules and lexicons created in consultation with radiology experts, the system then uses the information derived during the previous steps to perform NER for specific target entities.

EdIE-R has been shown to recognise named entities reliably accurately in brain imaging reports in the ESS data, which was used to write the original NER rules for this domain.[25] We have subsequently updated the rules based on the new development data from Tayside and all the results reported here are from the updated version. Performance on the ESS data has dropped very slightly from the earlier version but *EdIE-R* performs very well on the new data. The reliance on hand-crafted rules makes it potentially costly and time-consuming to adapt the system to a different dataset, for example, radiology reports for scans of other body parts or for other diseases as well as other types of raw text records such as pathology reports.

The initial *EdIE-R* rule writing was done iteratively in parallel with rounds of annotation done by domain experts before settling on an annotation scheme. Several rounds of annotation were carried out to create gold data for system development and evaluation (ESS, Tay and TayExt). Having this annotated gold data available provided us with the

[3]The tokenised and POS tagged output of EdIE-R was used to prepare the datasets used for evaluation by all systems described in this paper.

opportunity to try and test machine learning based methods which are typically used for NER on standard evaluation datasets (e.g. CoNLL or ACE data).

EdIE-N

EdIE-N represents a machine learning based approach to the problem of NER for radiology reports. As opposed to *EdIE-R*'s hand-crafted rules, *EdIE-N* infers named entity annotation from training data automatically, and applies these learned "rules" captured by the trained model to new data.

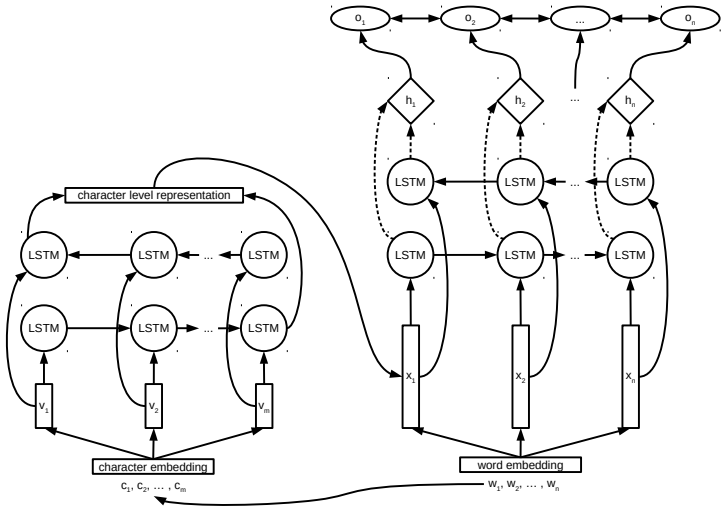

Figure 3: Schematic of *EdIE-N* entity recognition network.

In particular, the system makes use of *deep learning* via a neural network architecture (see Figure 3). *EdIE-N* employs a Conditional Random Field (CRF)[26] on top of a bi-directional LSTM[5] architecture to perform NER by assigning the best score s of consecutive of output labels o to a given sentence. The input features x_i of the classification network are comprised of word embeddings, which get concatenated with word representations derived from a character-level LSTM. This architecture is similar that of Huang et al. (2015)[27] and Zhou et al. (2015)[28] (see Figure 3).

Both word embeddings as well as character embeddings can either be learned during training from randomly initialized embedding matrices, or looked up in pre-trained models. At training time, the CRF output layer is conditioned on the LSTM hidden layer representations h_i. At test time, the system assigns entity annotations to each word according to the most likely entity type as determined by the CRF.

EdIE-N models can be trained either as a "monolithic" NER model, i.e. taking all possible entity types into account and potentially making use of interactions between them, or as a "bag-of-models" system, where one model is trained per unique entity type. We only report the results of the "bag-of-models" setup as using this approach makes it easier to add new entity types to an existing architecture. However, we have experimented with "monolythic" NER models which resulted in broadly similar performances to the latter.

As opposed to *EdIE-R*, the machine learning approach employed by *EdIE-N* does not rely on hand-crafted rules for NER. Instead, the system is trained on an annotated gold standard, from which entity type assignments are learned automatically. This alleviates the need for expert knowledge for designing new rules, making it both fast and inexpensive to learn and abstract from any given dataset. However, as a fully supervised machine learning approach, it does introduce the need for annotated gold data for training. Moreover, there is a common understanding that a sufficiently large training dataset is needed for a model to learn enough examples so that it performs reasonably well on new data. Creating such data is time-consuming. The other limitation to a machine learning based system is that it is very difficult to conduct error analysis and determine the exact reason for system errors.

SemEHR

The third approach we chose to compare to is a NER tool which was originally developed and trained for other purposes. The main goal is to compare the above two approaches with a generic portable tool that is able to be adapted for this particular stroke subtyping task. The tool picked for this purpose is *SemEHR*[29], which is an open source toolkit that integrates text mining and semantic computing for identifying mentions of UMLS[30] (Unified Medical Language System) concepts from clinical documents. Specifically, we adopted a *SemEHR* instance that has been trained on EHR data of South London and Maudsley, a psychiatric hospital in London. This instance was trained for identifying physical illnesses, such as liver diseases, HIV, diabetes etc. Each mention identified by *SemEHR* was associated with three-dimensional contextual information, i.e. negation (whether the condition was negated or affirmed), temporality (whether it was a recent or past event) and experiencer (whether the sufferer was the patient or other people).

SemEHR is based on GATE Bio-YODIE [4] and was adapted in two steps. The first step involved generating a mapping from what *SemEHR* identifies (i.e. mentions of UMLS concepts) to what this study is looking for (i.e. the entity types listed in Table 2). For those entity types not in UMLS vocabulary (e.g. *small vessel disease*), an additional dictionary is generated and combined with *SemEHR*'s existing gazetteer. There are cases where one UMLS concept is mapped to different entity types (e.g. *C0038454* is mapped to *stroke* and *ischaemic stroke*). To disambiguate them, the second adaptation step was to train a machine learning model on those cases. Details and source code of the second step are made available on GitHub.[31]

Experiments

We used strict CoNLL-style NER evaluation to compare the performance of the three systems on the different datasets and report individual scores per entity type and overall NER scores. We report precision (P), recall (R) and balanced F-score (F1), the harmonic mean of precision and recall. In the case of EdIE-N we average scores over 5 runs to account for fluctuations in classification results due to random initializations in the network models.

For EdIE-N first we report overall scores when training is performed on the development data of ESS, ESS plus Tay and all three development sets combined (ESS+Tay+TayExt). While the model trained on ESS data performs best on its own test data, we consider EdIE-N trained on all three datasets to be the better one as it performs best on the other two test sets and only slightly worse on ESS test (see Table 3). We use this model for the subsequent comparison.

Evaluation on Test data	ESS			Tay			TayExt		
Training data	P	R	F1	P	R	F1	P	R	F1
ESS	**0.85**	**0.93**	**0.89**	0.75	0.83	0.79	0.69	0.81	0.75
ESS+Tay	0.84	0.91	0.87	0.78	0.87	0.83	0.75	0.83	0.79
ESS+Tay+TayExt	0.82	0.92	0.86	**0.80**	**0.91**	**0.85**	**0.76**	**0.85**	**0.80**

Table 3: EdIE-N performance under different combinations of training and test data (best scores in **boldface**).

For our final comparison, we compare the rule-based *EdIE-R* system against the *EdIE-N* LSTM-CRF architecture and the *SemEHR* transfer learning approach on the ESS, Tayside, and extended Tayside test sets (see Tables 4, 5 and 6).

The hand-crafted rules of the *EdIE-R* system outperform the two machine learning approaches, even reaching near IAA levels on the ESS data. The gap between *EdIE-R* and *SemEHR*, the transfer learning approach incorporating rich information from out-of-domain resources, is relatively small, especially on the ESS and extended Tayside data. It is a little more pronounced on the original Tayside test set. Here, *EdIE-R* clearly benefits from rules specifically tailored to those two sets. Overall, *SemEHR* performs remarkably accurately on the test data, matching the hand-crafted system in recall, at the cost of losing some precision. On all datasets, the machine learning approach as addressed with *EdIE-N* falls behind the two other systems. However, these results are of little surprise, as the system uses no external knowledge such as access to an ontology, and relies entirely on features that are being derived automatically from the target texts. It is very likely that the performance of *EdIE-N* can be further improved by using additional training data, incorporating additional domain knowledge, or optimising model parameters further. While *EdIE-R* is the best overall system, there are certain labels, e.g. *subarachnoid haemorrhage*, for which *SemEHR* consistently performs better.

[4] https://gate.ac.uk/applications/bio-yodie.html

Evaluation on ESS test	EdIE-R			EdIE-N			SemEHR			IAA		
Entity Type	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ischaemic stroke	**1.00**	**1.00**	**1.00**	0.87	0.96	0.92	0.96	**1.00**	0.98	0.98	1.00	0.99
haemorrhagic stroke	0.80	0.84	0.82	0.74	0.76	0.75	**0.88**	**0.95**	**0.91**	0.93	0.99	0.96
stroke	**1.00**	**0.89**	**0.94**	0.26	0.35	0.30	0.92	**0.89**	0.90	1.00	0.96	0.98
glioma tumour	-	-	-	-	-	-	-	-	-	-	-	-
meningioma tumour	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	1.00	1.00
metastasis tumour	**1.00**	**1.00**	**1.00**	0.83	0.83	0.83	0.92	**1.00**	0.96	1.00	1.00	1.00
tumour	**0.98**	**0.99**	**0.99**	0.93	0.98	0.96	0.81	0.98	0.89	0.99	0.99	0.99
subdural haematoma	**0.73**	0.99	**0.84**	0.65	0.92	0.76	0.71	**1.00**	0.83	0.77	1.00	0.87
small vessel disease	0.93	**0.98**	0.95	0.57	0.85	0.69	**0.98**	0.95	**0.97**	0.95	0.97	0.96
atrophy	**0.96**	**0.98**	**0.97**	0.71	0.90	0.79	0.94	0.97	0.96	0.91	.96	.94
microhaemorrhage	**0.90**	**0.90**	**0.90**	0.00	0.00	0.00	0.83	0.50	0.63	1.00	1.00	1.00
subarachnoid haemorrhage	0.80	**0.80**	0.80	0.27	0.30	0.29	**1.00**	**0.80**	**0.89**	0.75	0.90	0.82
haemorrhagic transformation	0.25	**1.00**	0.40	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**	0.50	1.00	0.67
location:cortical	**0.98**	**0.98**	**0.98**	0.88	0.93	0.90	0.96	0.97	0.96	0.99	1.00	0.99
location:deep	**0.92**	0.88	0.90	0.90	**0.96**	**0.93**	**0.92**	0.91	0.91	0.95	0.94	0.94
time:old	**0.98**	**0.97**	**0.97**	0.94	0.97	0.95	0.92	0.88	0.90	0.97	0.98	0.98
time:recent	**1.00**	**1.00**	**1.00**	0.95	0.99	0.97	0.98	**1.00**	0.99	1.00	1.00	1.00
All	**0.94**	**0.96**	**0.95**	0.82	0.92	0.86	0.93	**0.96**	0.94	0.96	0.98	0.97

Table 4: NER results and IAA scores on the ESS test data.

Evaluation on Tay test	EdIE-R			EdIE-N			SemEHR			IAA		
Entity Type	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ischaemic stroke	**1.00**	**0.99**	**1.00**	0.78	0.97	0.86	0.99	**0.99**	0.99	0.90	1.00	0.96
haemorrhagic stroke	**0.97**	**0.93**	**0.95**	0.88	0.94	0.91	0.96	0.91	0.93	1.0	1.0	1.0
stroke	**1.00**	**1.00**	**1.00**	0.30	0.33	0.32	0.80	0.89	0.84	-	-	-
glioma tumour	0.80	0.44	0.57	0.00	0.00	0.00	**0.83**	**0.56**	**0.67**	-	-	-
meningioma tumour	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	-	-	-
metastasis tumour	**1.00**	**0.99**	**1.00**	0.88	0.94	0.91	0.91	**0.99**	0.95	1.0	1.0	1.0
tumour	**0.99**	**1.00**	**0.99**	0.88	0.95	0.91	0.97	0.58	0.72	0.89	1.00	0.94
subdural haematoma	**0.92**	**1.00**	**0.96**	0.70	0.89	0.78	0.87	0.84	0.86	1.00	1.00	1.00
small vessel disease	0.95	**0.95**	**0.95**	0.45	0.87	0.59	**0.96**	0.90	0.93	0.92	0.96	0.94
atrophy	**1.00**	**0.97**	**0.99**	0.81	0.93	0.86	**1.00**	**0.97**	0.98	1.00	1.00	1.00
microhaemorrhage	**1.00**	**0.67**	**0.80**	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
subarachnoid haemorrhage	**1.00**	0.69	0.82	0.35	0.38	0.36	**1.00**	**0.81**	**0.90**	0.80	1.00	0.89
haemorrhagic transformation	**1.00**	**1.00**	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-
location:cortical	**0.99**	**1.00**	**0.99**	0.95	0.96	0.96	0.98	0.99	**0.99**	0.96	1.00	0.99
location:deep	**1.00**	**0.82**	**0.90**	0.83	0.80	0.82	0.98	0.80	0.88	0.97	0.82	0.89
time:old	**0.98**	**0.98**	**0.98**	0.69	0.93	0.79	0.71	**0.98**	0.82	0.93	1.00	0.97
time:recent	**0.99**	**0.99**	**0.99**	0.92	0.99	0.95	0.98	**0.99**	0.98	0.98	1.00	0.99
All	**0.99**	**0.95**	**0.97**	0.80	0.91	0.85	0.96	0.89	0.91	0.95	0.96	0.96

Table 5: NER results and IAA scores on the Tay test dataset.

Conclusions

We have presented a system comparison for the task of labelling Named Entities in Electronic Health Records. Three approaches to the task were evaluated on three data sets. A hand-written system engineered by domain experts was able to consistently outperform a transfer learning system, applying previously established rules to new data, and a data-driven machine learning system.

The results confirm previously established findings that a hand-written, rule-based approach is able to perform NER on written EHR data very accurately, albeit for a high development cost in terms of time and effort afforded by domain experts. While the machine learning approach performed worse in our comparison, we are still slightly optimistic that such an approach can be reasonably employed where there are either no experts readily available, or a system has to

Evaluation on TayExt test	EdIE-R			EdIE-N			SemEHR		
Entity Type	P	R	F1	P	R	F1	P	R	F1
ischaemic stroke	**0.98**	**0.93**	**0.95**	0.75	0.91	0.82	0.95	0.85	0.90
haemorrhagic stroke	**0.86**	0.74	0.79	0.68	0.64	0.66	0.85	**0.80**	**0.82**
stroke	**1.00**	0.60	0.75	0.25	0.40	0.31	0.80	**0.80**	**0.80**
glioma tumour	**0.86**	**1.00**	**0.92**	0.00	0.00	0.00	**0.86**	0.50	0.63
meningioma tumour	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
metastasis tumour	**1.00**	**1.00**	**1.00**	0.83	0.86	0.85	**1.00**	**1.00**	**1.00**
tumour	**0.95**	0.86	**0.91**	0.73	**0.92**	0.82	0.81	0.82	0.81
subdural haematoma	**0.86**	**0.87**	**0.86**	0.61	0.73	0.66	0.80	0.85	0.83
small vessel disease	**0.95**	**0.84**	**0.89**	0.36	0.72	0.48	0.89	0.75	0.82
atrophy	**1.00**	0.94	0.97	0.89	0.94	0.92	0.99	**0.99**	**0.99**
microhaemorrhage	**1.00**	**0.50**	**0.67**	0.00	0.00	0.00	0.00	0.00	0.00
subarachnoid haemorrhage	0.94	0.63	0.76	0.36	0.35	0.36	**0.98**	**0.83**	**0.90**
haemorrhagic transformation	0.28	**0.70**	0.40	0.00	0.00	0.00	**0.36**	0.50	**0.42**
location:cortical	**0.98**	**0.99**	**0.99**	0.94	0.97	0.95	0.97	**0.99**	0.98
location:deep	0.91	**0.96**	0.93	0.77	0.87	0.81	**0.95**	0.91	**0.94**
time:old	**1.00**	**0.93**	**0.96**	0.78	0.92	0.84	0.82	0.90	0.86
time:recent	**1.00**	0.95	0.97	0.90	0.93	0.92	0.97	**0.99**	**0.98**
All	**0.94**	**0.91**	**0.93**	0.76	0.85	0.80	0.91	**0.91**	0.91

Table 6: NER results on the TayExt test dataset.

be developed quickly for a relatively low cost. The transfer learning approach showed impressive results, presenting a viable alternative to an entirely hand-written system, though still requiring a good deal of human manipulation.

In the future, we would like to further improve the ways to reliably and automatically label Named Entities in EHRs. Of particular interest are more experiments on the machine learning approach, where especially more fine-grained tuning of hyper parameters promises to be able to yield better performance results. Additionally, we would like to explore the possibility of combining the approaches presented in this paper. The modular nature of the overall EHR processing pipeline could enable us to employ the different NER systems according to their individual strengths and weaknesses, potentially leading to a better overall performance downstream, e.g, at the document labelling stage. Another interesting future direction is the rapid development of new systems in multiple iterations. When required, one could start by rapidly and inexpensively adding a new label to the system using the machine learning approach, and subsequently improving on it by utilising its results to guide the transfer or hand-crafting of reliable rules.

Acknowledgements

Gorinski, Tobin, Grover, Alex and Whalley are supported by the MRC Mental Health Data Pathfinder Award (MRC - MC_PC_17209). Wu is MRC/Rutherford fellow of HRD UK (MR/S004149/1). Grover was and Alex is supported by The Alan Turing Institute (EPSRC grant EP/N510129/1). Whiteley was supported by an MRC Clinician Scientist Award (G0902303) and is supported by a Scottish Senior Clinical Fellowship (CAF/17/01). Sudlow is Chief Scientist of UK Biobank and Director of HDR UK Scotland.

References

1. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007, 30(1):3-26.
2. Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing, 2011, pp. 1524-34.
3. Rocktäschel T, Weidlich M, Leser U. ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics. 2012, 28(12):1633-40.
4. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. In: Biocomputing 2008, pp. 652-63.
5. Schuster M, Kuldip KP. Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing,1997, pp. 2673-81.
6. Cornegruta S, Bakewell R, Withey S, Montana G. Modelling radiological language with bidirectional long short-term memory networks. In: Proceedings of the 7th International Workshop on Health Text Mining and Information Analysis, 2016, pp. 17-27.

7. Tjong KS, Erik F, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In: Proceedings of CoNLL, 2003, pp. 142-7.
8. Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, 2009, pp. 147-55.
9. McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003, pp. 188-91.
10. Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 2005, pp. 363-70.
11. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002, pp. 168-75.
12. Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications 2004 Aug 28, pp. 104-7.
13. Ogren PV, Savova GK, Chute CG. Constructing evaluation corpora for automated clinical named entity recognition. InMedinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems 2007, IOS Press, p. 2325.
14. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 Sep 1, 17(5):507-13.
15. Alex B, Haddow B, Grover C. Recognising nested named entities in biomedical text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. 2007, pp. 65-72.
16. Grover C, Haddow B, Klein E, Matthews M, Nielsen LA, Tobin R, Wang X. Adapting a relation extraction pipeline for the BioCreAtIvE II task. Proceedings of the BioCreAtIvE II Workshop. 2007, pp. 273-86.
17. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of Medical Informatics. 2008(01);17:128-44.
18. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016(2);279:329-343.
19. Flynn RWV, Macdonald TM, Schembri N, Murray GD, Doney ASF. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. Pharmacoepidemiology and drug safety. 2010, 19(8):843-7.
20. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 2010 Oct 1, 22(10):1345-59.
21. Nothman J, Ringland N, Radford W, Murphy T, Curran JR. Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence. 2013 Jan 1, 194:151-75.
22. Arnold A, Nallapati R, Cohen WW. Exploiting feature hierarchy for transfer learning in named entity recognition. In: Proceedings of ACL-08: HLT. 2008:245-53.
23. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160-7.
24. Jackson C, Crossland L, Dennis M, Wardlaw J, Sudlow C. Assessing the impact of the requirement for explicit consent in a hospital-based stroke study. QJM: Monthly Journal of the Association of Physicians. 2008, 101(4):281-9.
25. Alex B, Grover C, Tobin R, Sudlow C, Mair G, Whiteley W. Text Mining Brain Imaging Reports. Journal of Biomedical Semantics, accepted for a special issue to appear in 2019, preprint available at `https://www.research.ed.ac.uk/portal/files/77559477/Text_Mining_Brain_Imaging_Reports.pdf`.
26. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, 2001, pp. 282-9.
27. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991, 2015.
28. Zhou J, Xu W. End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015, pp. 1127-37,
29. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, Kartoglu I, Agrawal A, Stringer C, Gale D, Gorrell G, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. Journal of the American Medical Informatics Association, 2018, 25(5):530-7.
30. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 2004, 32:D267–D270.
31. `https://github.com/CogStack/nlp2phenome`