# Not a cute stroke: Analysis of Rule- and Neural Network-Based Information Extraction Systems for Brain Radiology Reports

**Andreas Grivas**[†] **Beatrice Alex**[‡†] **Claire Grover**[†] **Richard Tobin**[†] **William Whiteley** [§]

[†]School of Informatics, [‡]Edinburgh Futures Institute, [§]Centre for Clinical Brain Sciences, [§]Usher Institute
University of Edinburgh, United Kingdom

[§]Nuffield Department of Population Health
University of Oxford, United Kingdom
{agrivas|balex|grover|richard}@inf.ed.ac.uk

## Abstract

We present an in-depth comparison of three clinical information extraction (IE) systems designed to perform entity recognition and negation detection on brain imaging reports: EdIE-R, a bespoke rule-based system, and two neural network models, EdIE-BiLSTM and EdIE-BERT, both multi-task learning models with a BiLSTM and BERT encoder respectively. We compare our models both on an in-sample and an out-of-sample dataset containing mentions of stroke findings and draw on our error analysis to suggest improvements for effective annotation when building clinical NLP models for a new domain. Our analysis finds that our rule-based system outperforms the neural models on both datasets and seems to generalise to the out-of-sample dataset. On the other hand, the neural models do not generalise negation to the out-of-sample dataset, despite metrics on the in-sample dataset suggesting otherwise.

## 1 Introduction

Information Extraction (IE) from radiology reports is of great interest to clinicians given its potential for automating large scale data linkage, targeted cohort selection, retrospective statistical analyses, and clinical decision support (Pons et al., 2016). Accurate IE from radiology reports has also received a surge of attention due to the insatiable demand of deep learning medical image classifiers for more labelled training data (Irvin et al., 2019).

While IE from radiology reports is of increasing value, the scarcity of annotated data and limited transferability of previously developed models is currently hindering progress. Despite recent breakthroughs in learning contextual representations for clinical and biomedical text from large amounts of unlabelled text (Devlin et al., 2019; Peng et al., 2019; Alsentzer et al., 2019; Lee et al., 2019), labelled data scarcity remains the bottleneck to improvements and wider adoption of deep learning methods. Data scarcity is even more prominent in the general clinical domain with its vast quantity of possible entity labels.

Existing approaches to overcome the lack of labelled data include using a rule-based system to annotate more data (Smit et al., 2020) or propose labels in an annotation tool (Nandhakumar et al., 2017; Alex et al., 2019; Searle et al., 2019), leveraging semi-supervised learning to speed up annotation (Wood et al., 2020) and creating artificial data (Schrempf et al., 2020). It is also common for rule-based systems to be developed alongside statistical models to contrast their performance (Cornegruta et al., 2016; Gorinski et al., 2019; Sykes et al., 2020). We need to understand the shortcomings and benefits of rule-based and neural models to improve annotation decisions and system evaluation, a comparison which we explore in this paper both on in- and out-of-sample data.

The use of end-to-end learning for document labelling has been a recent trend in analysing radiology reports (Smit et al., 2020; Schrempf et al., 2020). Contextual representations of a document such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) are used as input to a multi-label classifier to label the report directly without first recognising named entities and negation. While such approaches make annotation simpler and faster and rely less on complex modelling decisions, they have various shortcomings. Firstly, they lack in interpretability, as it is hard to probe which parts of a document the model uses when making predictions. Some models employ an attention mechanism highlighting tokens in the input used to arrive at the decision (Mullenbach et al., 2018; Schrempf et al., 2020). However, they are opaque as to the exact sub-decisions that lead to the labels, which is unsatisfactory in the clinical domain where interpretability is of paramount importance. Secondly,



*Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 24–37

https://doi.org/10.18653/v1/P17

they are not data-efficient. For example, Smit et al. (2020) predict four labels per entity type (positive, negated, uncertain and blank). To scale such approaches to more entity types a lot of annotated data is needed, the absence of which is currently a limiting factor. Lastly, a significant drawback of end-to-end approaches is that no part of the system other than the encoder is reusable in any domain that has a different non-overlapping set of output labels. For that new domain, the labelling procedure needs to be initiated from scratch, leading to a duplication of effort.

In this work, as in some previous neural approaches (Bhatia et al., 2019) and as is common in rule-based approaches (Cornegruta et al., 2016; Fu et al., 2019), we employ a bottom up approach to document labelling by factoring the problem into sub-tasks. This way document labels are interpretable as a sequence of decisions with some sub-tasks being extendable and reusable on other datasets. Our three IE systems, EdIE-R, EdIE-BiLSTM and EdIE-BERT (a rule-based and two neural models), recognise mentions of stroke, stroke sub-types and other related *findings* such as tumours and small vessel disease in text. They also identify related temporal *modifiers* (recent or old) and location *modifiers* (deep or cortical). For downstream document classification by phenotype, the systems also mark findings and modifier entities for negation (*negation detection*).

The contributions of our work are three-fold:

1. We compare our systems both on an in-sample and an out-of-sample dataset, drawing attention to generalisation issues of our neural models' negation detection on the out-of-sample dataset which are opaque when inspecting metrics on the in-sample one.
2. We draw on our error analysis to highlight ways in which using previously developed systems to suggest labels for new data can go wrong and propose using pretrained neural contextual models, such as BERT, to detect and correct inconsistencies.
3. We make our code[1], models and web interface[2] publicly available for re-use on brain imaging reports, as a way to bring the software to the data and assist research in this area.

[1] https://github.com/Edinburgh-LTG/edieviz
[2] http://jekyll.inf.ed.ac.uk/edieviz/

## 2 Related Work

**Named entity recognition** (NER) is a standard natural language processing (NLP) task and is commonly limited to identifying proper nouns in text (e.g., *person*, *organisation*, and *location*) (Sang and Meulder, 2003). In the clinical domain concepts of interest are usually *problems*, *tests* and *treatments*, as formulated in the clinical concept extraction i2b2 shared task (Uzuner et al., 2011). In our case, as in previous work on text mining and IE applied to radiology reports (Hassanpour and Langlotz, 2016; Cornegruta et al., 2016; Zhu et al., 2019), we use NER to refer to recognising entities that are either relevant medical *findings*, such as *ischemic stroke*, or *modifiers*, such as *acute*.

Approaches for NER in this domain, while not mutually exclusive, can broadly be categorised into the following: approaches leveraging lexicons, such as cTAKES (Savova et al., 2010) and RadLex (Langlotz, 2006); ontologies, such as MetaMap (Aronson and Lang, 2010); rule-based systems and pattern matching (Cornegruta et al., 2016); feature based machine learning such as Conditional Random Fields (CRFs) (Hassanpour and Langlotz, 2016); and more recently, deep learning (Cornegruta et al., 2016; Zhu et al., 2019).

**Negation detection** is commonly framed as identifying negation or speculation cues and their matching scopes in sentences (Fancellu et al., 2017). In the clinical domain, however, it is common for approaches to tackle negation assertion, namely, to verify whether each identified entity mention in the text is negated or affirmed (Bhatia et al., 2019), and in some cases, whether it is uncertain (Peng et al., 2018), conditionally present, hypothetically present or relating to some other patient (Uzuner et al., 2011).

As with NER, some of the earlier negation detection approaches were rule-based. NegEx (Chapman et al., 2001) relies on regular expressions to detect negation patterns, and has been successfully applied to discharge summaries. Hassanpour and Langlotz (2016) and Cornegruta et al. (2016) use NegEx for negation detection on extracted entities.

Context (Harkema et al., 2009) extends NegEx to capture hypothetical mentions, experiencer information and temporality, albeit with limited success on the latter. NegBIO (Peng et al., 2018), another rule-based negation and uncertainty detection system extended through dependency parsing information, has been shown to outperform NegEx.

Similarly, Cornegruta et al. (2016) demonstrated that enhancing NegEx with Stanford dependencies outperformed their bidirectional LSTM (BiLSTM) negation model. BiLSTM approaches for negation detection have been successful, with Fancellu et al. (2017) reporting state of the art results for BioScope (Vincze et al., 2008) abstracts. Sergeeva et al. (2019) outperformed the latter using pretrained transformer models.

Despite the amount of progress on negation detection for clinical texts, however, there is still ample evidence that while fitting systems on a particular dataset is straightforward, generalising negation detection across datasets is challenging (Wu et al., 2014). This is true both for out-of-domain evaluation, such as training on a dataset of medical articles with evaluation on a dataset of clinical text (Wu et al., 2014; Miller et al., 2017), as well as for out-of-sample evaluation, where the training and test datasets are from the same domain but may have differences due to different annotation style, or distribution of named entities (Sykes et al., 2020). For the in-domain but out-of-sample case, a domain fine-tuned rule based system seems to transfer well (Sykes et al., 2020). For all other cases, transfer is challenging, both for rule-based and machine-learning models (Wu et al., 2014; Miller et al., 2017; Sykes et al., 2020), with machine learning models benefiting from the addition of in-domain data to the training set. Lin et al. (2020) demonstrate that a pretrained BERT model can improve the results of domain transfer for negation detection, but the results are still lower for out-of-domain datasets than in-domain datasets if we compare to the results of earlier models in Miller et al. (2017). In our work we concur with previous findings: our neural models do not generalise negation detection across datasets, despite both datasets comprising radiology reports with stroke findings, such as acute ischemic stroke (AIS).

**Document classification** In our work, we formulate NER and negation detection as sub-tasks towards document classification by phenotypes and will report derived document classification results for one label (acute ischemic stroke) on a freely available data set of brain MRI radiology reports (Kim et al., 2019) with the aim of testing generalisability of our systems. Kim et al. (2019) compared different machine learning approaches on this data and found a single decision tree performed best (precision=91.1, recall=95.3, F1=93.2 and acc=98%) on labelling reports as AIS or non-AIS phenotypes.

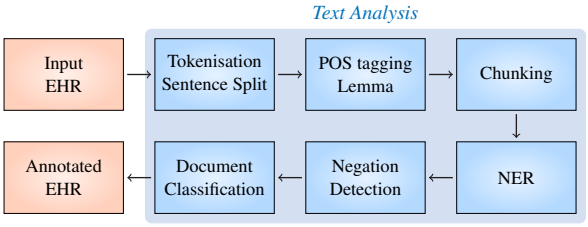


Figure 1: EdIE-R pipeline.

On a dataset in a similar domain, Fu et al. (2019) classify a set of 1000 radiology reports by Silent Brain Infarct. Their rule-based system based on MedTagger is superior to their word-level Convolutional Neural Network (CNN) for predicting Silent Brain Infarcts, but not for White Matter Disease.

For a broader exposition of NLP applied to radiology reports, we refer to Pons et al. (2016).

## 3 System Descriptions

The rule-based system, EdIE-R, and the neural systems, EdIE-BiLSTM and EdIE-BERT, all factor the document labelling task into the same three subtasks. Namely, extracting finding mentions, extracting modifier mentions and using negation detection to assert whether the mentions imply their presence or absence in the brain imaging report. All three systems work at a sentence level granularity.

### 3.1 EdIE-R

The rule-based system consists of a pipeline with four main components which are applied in sequence (see Figure 1). Two components perform linguistic analysis of the text of radiology reports, namely, NER for finding and modifier predictions and negation detection to distinguish between affirmative and negative instances. The third component computes document-level labels based on the preceding linguistic analysis. These main components are preceded by text pre-processing steps, i.e. tokenisation, part-of-speech tagging (POS) and shallow chunking.

The EdIE-R components make use of handcrafted rules and lexicons which were created in consultation with radiology experts. The rules and lexicons are applied using the XML tools LT-XML2 (Grover and Tobin, 2006), in combination with Unix shell scripting. The NER rules are lexicon- and regular expression-dependent but the quality of the POS tagging and lemmatisation is also important. We use the C&C POS tagger (Cur-

ran and Clark, 2003) with a standard model trained on newspaper text as well as a model trained on the Genia biomedical corpus (Kim et al., 2003). After running the POS tagger with each of the models, we apply a rule-based correction stage to moderate disagreements between them. After POS tagging, we apply the *morpha* lemmatiser (Minnen et al., 2000) to analyse inflected nouns and verbs and compute their lemmas. The negation detection component relies partly on the pre-processing (i.e. recognition of negation-bearing tokens such as *no*, *not*, *n't*), and partly on the output of the chunker, which is used to constrain the scope of negative particles. The neural models we introduce in the next section rely on EdIE-R's preprocessing pipeline.

### 3.2 EdIE-BiLSTM

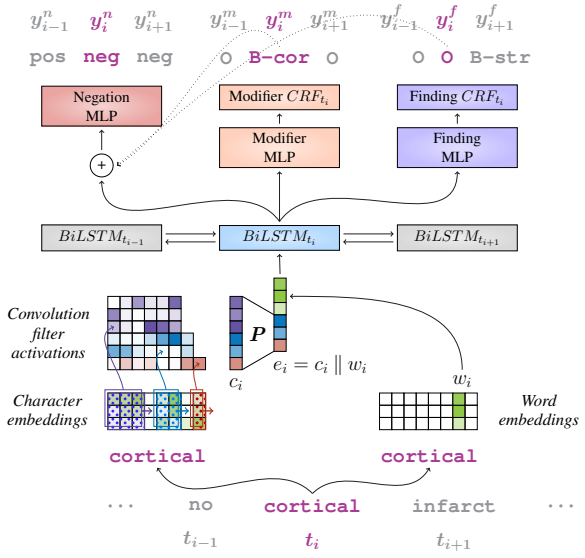


Figure 2: EdIE-BiLSTM model with multitask output for negation, finding and modifier prediction. Current input $t_i$ and outputs $y_i^{task}$, such as B-cortical, are highlighted. Other timesteps appear in grey.

In order to predict the spans of findings and modifiers jointly with whether they are negated, we frame these sub-tasks as an instance of multitask learning (Caruana, 1997), similar to Bhatia et al. (2019), and train a neural network model. The network depicted in Figure 2 has three task heads, with a Conditional Random Field (CRF) output for modifiers and findings and a sigmoid binary classifier for negation. To condition negation on finding and modifier predictions, we feed the predicted findings and modifiers to the negation Multilayer Perceptron (MLP) by adding a learned embedding to the activations of the tokens that have been tagged as findings or modifiers. We do not encode entity type to avoid biasing our negation detector towards using type:negation correlations, since as we shall see, such biases do not transfer across datasets. We decide negation for tagged entities by assigning the negation prediction made for the entity's first token. The part of the architecture mentioned so far is the same for both EdIE-BiLSTM and EdIE-BERT, the two models differ solely as to their choice of sentence encoder.

The encoder for EdIE-BiLSTM is a character CNN - word BiLSTM with randomly initialised embeddings. Given such an initialisation, it has no preconceptions about the text in the dataset and can flexibly fit the data, with the risk of overfitting. We obtain character aware token embeddings $c_i$ by using a character level convolutional network following a modified version of the small CNN encoder model of Kim et al. (2016) (see Appendix 3.2 for details). A word-level embedding is obtained by concatenating a projected character-level token representation with a word embedding $e_i = c_i \parallel w_i$. Context-aware representations for sentence tokens are computed by propagating word-level representations through a BiLSTM network.

### 3.3 EdIE-BERT

EdIE-BERT only differs from EdIE-BiLSTM by replacing the BiLSTM encoder with a pretrained BERT (Devlin et al., 2019) encoder. More specifically, since we are working with radiology reports, we elected to use BlueBERT (Peng et al., 2019), which is a BERT model that is adapted for the clinical domain by being pretrained further on PubMed biomedical abstracts and clinical texts from the MIMIC-III dataset (Johnson et al., 2016). While there is a menagerie of similar models to BlueBERT, such as BioBERT (Lee et al., 2019) and ClinicalBERT (Alsentzer et al., 2019), BlueBERT was found to outperform them when used for radiology report document classification (Smit et al., 2020). As a pretrained alternative, the BlueBERT encoder comes with preconceptions about clinical text. For example, synonyms occurring in similar contexts are likely to have similar representations and hence be assigned similar predictions by the classification layer. We shall see that this results in EdIE-BERT having increased recall but also many false positives because of flagging similar concepts that were not annotated in the data.

## 4 In-Sample Evaluation

In this section we describe the dataset we used to develop and fit our systems and report in-sample performance of our models on the unseen test set.

| | ESS Dev | ESS Test |
|---|---|---|
| Reports | 364 | 266 |
| Sentences | 3,837 | 2,855 |
| Tokens | 32,229 | 22,842 |
| Findings | 2,373 | 1,494 |
| Modifiers | 1,959 | 1,430 |
| Total Entities | 4,332 | 2,924 |

Table 1: ESS data statistics.

### 4.1 Edinburgh Stroke Study (ESS) Dataset

The ESS dataset (Alex et al., 2019) is comprised of English text reports produced by radiologists which describe findings in imaging reports. The reports are predominantly computerised tomography (CT) brain imaging reports with fewer magnetic resonance imaging (MRI) scans collected for a regional stroke study.

An example of a radiology report can be seen in Figure 7 in the Appendix. The language of radiology reports is usually short and descriptive as it is limited to descriptions of the image. Negation is usually overt (Sykes et al., 2020), e.g. "*no visible infarct*" with occasional hedging, e.g. "*there may be early signs of deterioration*". There is some variation in radiologist styles, some use note style, others use full sentences.

Manual annotation of the reports was accomplished in tranches by two experts, a neurologist and a radiologist, correcting output of an early version of EdIE-R. The data was split into development (dev) and test data (see Table 1). Annotations include different entity types (12 finding types and 4 modifier types), relations between corresponding modifier and finding entities, negation of entities, and 24 document level labels (phenotypes). We note that negation labels are binary and are only assigned to findings or modifier entities. Annotators were instructed to mark any mention of findings and modifiers not clearly indicated to be present as negated. This paper focuses only on the entity and negation annotation. The entire ESS test set was doubly annotated to allow us to calculate inter-annotator agreement (IAA) using precision, recall and F1. IAA F1 is 96.15 for findings and 97.83 for modifiers. The combined NER and negation IAA F1 is 96.11.

The version of EdIE-R presented here was further optimised on ESS dev. EdIE-BiLSTM and EdIE-BERT were trained using 285/364 (≈80%) of ESS dev reports, with validation and hyperparameter tuning performed on the remaining 79 (≈20%) reports. We report results on the unseen ESS test set which was not used for system development and hyperparameter tuning.

| System | Task | P | R | F1 |
|---|---|---|---|---|
| EdIE-R | Mod | **97.23** | 95.73 | **96.48** |
| | Find | **90.67** | **95.58** | **93.06** |
| | Neg | **92.46** | 94.32 | **93.38** |
| EdIE-BiLSTM | Mod | 94.99 | 95.38 | 95.18 |
| | Find | 90.54 | 94.85 | 92.64 |
| | Neg | 91.04 | 93.43 | 92.22 |
| EdIE-BERT | Mod | 94.66 | **96.71** | 95.68 |
| | Find | 86.06 | 95.05 | 90.33 |
| | Neg | 88.94 | **94.63** | 91.70 |

Table 2: Results for predicting finding (Find) and modifier (Mod) entities as well as their negation (Neg) in the ESS test set. Best system per task in bold.

We used CoNLL scoring which considers a system annotation as true positive only if both the entity span and the label are correct as represented in IOB encoding (Sang and Meulder, 2003). Negation detection F1-score is computed for the predicted findings and modifiers, and hence includes error propagation from those tasks. F1-scores are computed using precision (P) and recall (R) based on the number of true positives (TP), false positives (FP) and false negatives (FN).

### 4.2 Results

The performance of all models is high, but EdIE-R outperforms both neural models in precision and F1-score on all sub-tasks (Table 2). EdIE-BiLSTM outperforms EdIE-BERT in F-score at detecting findings. We hypothesise this is because randomly initialised embeddings have little prior bias and can fit any potential annotation inconsistencies unhindered. Lastly, EdIE-BERT has lower precision but high recall, which suggests that the model overzealously flags plausible spans as findings.

Our results so far seem to suggest that EdIE-BERT is the worst performing model overall for detecting findings. This comes as a surprise, since other models using BlueBERT have reported state of the art results on many tasks (Peng et al., 2019; Smit et al., 2020). However, as we shall see in the error analysis of Section 6, its errors are mostly false positives and span-mismatch errors. When looking at EdIE-BERT output, a large part of the errors are plausible and may be spans that were

missed or have boundaries that were annotated inconsistently.

We also note that both neural models underperform EdIE-R on negation, but not by a lot. They seem to generalise sensibly to the test set based on the $\approx 92$ F1 score, but as we shall see in the next section, this in-sample high score is misleading.

## 5 Out-of-Sample Evaluation

We now test how well each of our systems generalises to unseen radiology reports from a different source, highlighting that our neural models do not generalise negation detection to this other dataset. By out-of-sample, we mean this dataset has similar labels but comes from a different distribution than the one the systems were developed on. We emphasise that we have not trained or adapted our models to this dataset.

### 5.1 AIS Dataset

We evaluate all three systems on a dataset of brain MRI reports labelled for AIS, collected at Hallym University Chuncheon Sacred Heart Hospital in South Korea and made publicly available in Kim et al. (2019). The data is labelled with binary AIS labels at the report level which correspond to the presence or absence of AIS in the report.

The data contains reports for 432 patients with MRI readings of confirmed AIS. To create it, a neuroradiologist read MRI images, and the labelling of the corresponding reports as AIS or non-AIS was derived from these readings. The 2,592 non-AIS reports are from patients who underwent MRI brain imaging for a variety of reasons not related to ischaemic stroke. Kim et al.'s training set (70%) contains 303 AIS and 1,815 non-AIS reports, and their test set (30%) contains 129 AIS and 777 non-AIS reports. We note that the non-AIS reports are from MRI scans that were carried out for non-stroke related reasons, which likely makes this task much easier than in the general setting. Since the data is shared as one file containing all reports without specifying the exact split, we used the combined train and test data for our experiment (see Table 3).

When testing on the AIS data we compute precision, recall and F1 (and other metrics reported by Kim et al., 2019) but, in contrast to the ESS data, we are dealing with document label predictions. We inferred AIS and non-AIS labels based on whether there was a sentence in the report which contained both an ischaemic stroke finding and an acute modifier (AIS), or not (non-AIS). Our temporal modifier *time recent* overlaps well with the use of "*acute*" in the AIS data, with the exception of the term "*sub-acute*". For the purpose of inferring AIS labels, we therefore defined the acute modifier accordingly by excluding sub-acute mentions.

| | AIS Train & Test Data |
|---|---|
| Reports | 3,024 |
| Sentences | 22,280 |
| Tokens | 168,718 |
| AIS labelled reports | 432 |
| non-AIS labelled reports | 2,592 |

Table 3: AIS data statistics. The sentence and token figures are determined using the EdIE-R tokenisation and sentence detection.

| System | SP | NPV | P | R | F1 | Acc (%) |
|---|---|---|---|---|---|---|
| EdIE-R | 96.64 | **99.56** | 82.55 | **97.45** | **89.38** | **96.70** |
| EdIE-BiLSTM | **97.92** | 93.65 | **82.80** | 60.19 | 69.71 | 92.53 |
| EdIE-BERT | 97.18 | 94.56 | 79.72 | 66.44 | 72.47 | 92.79 |
| *Kim et. al 2019* on 30% of the data | 98.50 | 99.20 | 91.10 | 95.30 | 93.20 | 98.00 |

Table 4: Results for classifying reports as AIS. We report the same metrics as Kim et al. 2019 but on all of the AIS data: Specificity (SP), Negative Predictive Value (NPV), Precision (P≡Positive Predictive Value), Recall (R≡Sensitivity), F1-score and Accuracy (Acc).

### 5.2 Results

Table 4 shows that EdIE-R achieves an F1-score of 89.38. The results are lower than the best results reported in Kim et al. (2019), but this is partly to be expected since we do not adapt any of our systems to the AIS dataset, apart from formulating the document level rules. Interestingly, EdIE-R's recall was two points higher but its precision was considerably lower. A neurologist examined some of the false positives which contributed to EdIE-R's lower precision and reported that they did, in fact, indicate acute ischaemic stroke. It is possible that in these cases AIS was not the primary finding and that these reports were therefore not labelled as AIS. Given that our systems are configured to recognise all findings in a report at the entity level, it is not surprising to find a difference in predictions as compared to a binary document labelling system, but we consider the EdIE-R results to be an effective validation of our approach and can show that it generalises to other similar data.

Both neural systems had much worse results than EdIE-R, mostly due to considerably lower recall, demonstrating poor generalisation. On inspection of their predictions, we found that this

was overwhelmingly due to errors in negation detection. When removing negation, the results were very similar to those of EdIE-R. We found that one reason for the discrepancy is that the distribution of negation over findings in the ESS dataset compared to AIS is very different, with acute ischemic stroke being negated much more often in the ESS dataset compared to the AIS dataset. Despite not providing finding and modifier information to the negation detection head explicitly, the neural models seem to be using superficial features such as the distribution of negation for *acute* and *ischemic stroke* rather than relying on other features, such as overt negation cues, that would generalise. In this respect, our findings are similar to Fancellu et al. (2017), who demonstrated that neural network models were using punctuation as a cue for negation scope detection and failing to generalise beyond that.

## 6 Error Analysis

In this section we provide a fine-grained breakdown of the types of errors made by EdIE-R, EdIE-BiLSTM and EdIE-BERT, arguing that not all error types are equally detrimental to the downstream task of document labelling. Next, we investigate the variability in error types between our systems by exploiting BlueBERT's context-aware embeddings to group together training and evaluation examples that are similar. We then compare their labels to identify annotation artefacts that influence system errors. Lastly, we investigate how our systems handle spelling errors.

### 6.1 Breakdown of Error Types

As alluded to in Section 4.2, CoNLL F1 score harshly penalises wrong entity boundaries by reducing both precision and recall simultaneously (Finkel et al., 2005). For a deeper understanding of the situation, we dissect the errors (FP and FN counts) on the ESS test set into the following types (Manning, 2006):

*False Positive (FP)*: predicted spurious entity
*False Negative (FN)*: missed gold entity
*Label Error (LE)*: correct span, wrong label
*Boundary Error (BE)*: span overlap, correct label
*Label & Boundary Error (LBE)*: span overlap + LE

We note that when relying on finding and modifier predictions for document classification by phenotype, some errors are worse than others. We disfavour FPs, LEs and LBEs as they are likely to deteriorate document classification by hallucinating phenotypes. We also dislike FNs, but less so, since usually radiology reports have some degree of redundancy. Lastly, we argue the BEs are mostly benign, since for document classification the span of an entity should not affect label allocation.

| System | Task | FP | FN | LE | BE | LBE |
|---|---|---|---|---|---|---|
| EdIE-R | Mod | 33 | 50 | 0 | 1 | 5 |
| | Find | 100 | 20 | 1 | 37 | 7 |
| EdIE-BiLSTM | Mod | 45 | 37 | 3 | 13 | 5 |
| | Find | 96 | 27 | 10 | 34 | 5 |
| EdIE-BERT | Mod | 50 | 17 | 1 | 16 | 5 |
| | Find | 164 | 13 | 11 | 45 | 4 |

Table 5: Number of error types made by each system for findings and modifiers in ESS test set.

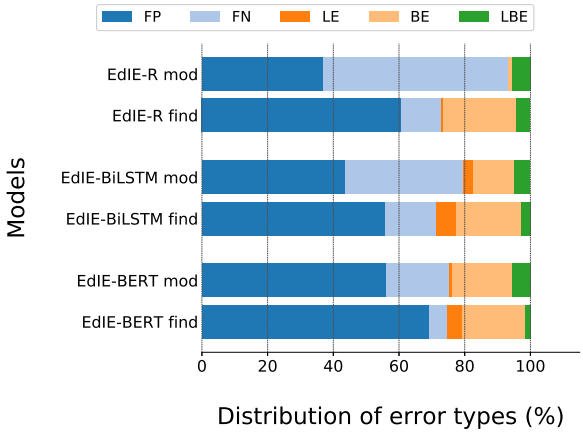


Figure 3: Proportion of error types made by each system for findings and modifiers in ESS test set.

Table 5 and Figure 3 show that EdIE-BiLSTM and EdIE-BERT make a larger proportion of LEs, which we found to be mostly due to ambiguity in annotation between *haemorrhagic stroke* and *stroke*. There was only one BE by EdIE-R, in contrast to more than ten by EdIE-BiLSTM (13) and EdIE-BERT (16), but on recognising findings, all three systems make more than 18% BEs. The large percentage of BEs, especially on modifiers, suggest inconsistencies in span selection during annotation. Such span inconsistencies unfairly lower the score of models when evaluating by NER F1. We conclude that care is needed when relying on a subtask metric that may not correlate with the document labelling goal as well as initially expected.

A striking difference is that EdIE-BERT has a larger proportion of FPs than the other systems, with the remaining errors being mostly BEs. This highlights that the model flags multiple spurious spans that are not annotated in the data, which as we shall see in the next section, is mostly due to

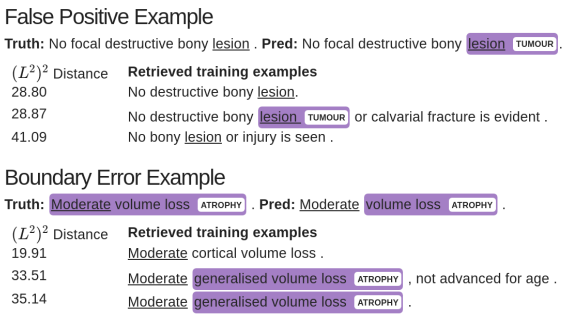


Figure 4: A EdIE-BERT *False Positive* and *Boundary Error* from the ESS test set. **Truth** and **Pred** are the gold annotations and EdIE-BERT predictions respectively. The sentences below are the three most similar training examples ordered by decreasing similarity.

inconsistencies in annotation than to model errors.

Lastly, EdIE-BiLSTM's larger proportion of FNs can be partly attributed to abbreviations, since EdIE-BiLSTM misses some not present in the training set, such as *CADASIL* (Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy) and *PICH* (Primary Intracerebral Haemorrhage). On the other hand, interestingly, EdIE-BERT tags some abbreviations that were unseen during training, such as *METS*, which can be short for *metastatic tumour*.

### 6.2 Nearest Neighbour Annotations

In this section, we exploit a pretrained[3] BlueBERT model's context-aware embeddings to group together sentence examples from all ESS data that are similar. We do so to gain insight into any potential annotation artefacts by contrasting the annotations of similar examples.

We follow Khandelwal et al. (2020): see equations (1) and (2) in their paper for technical details. We create a datastore with key value pairs, where the keys are BlueBERT embeddings of each token in the ESS training set and values are the token's labels. We then conduct an error analysis. For each token EdIE-BERT mislabelled during evaluation, we find the $k$ nearest neighbour tokens from the training set and visualise their labels.

In Figure 4 we plot two examples of EdIE-BERT errors on *findings*, a FP and a BE. Above on the left is the gold annotation with the prediction on the right and the error underlined. Below are the three most similar training examples as ordered by decreasing similarity using BlueBERT[4]. In the FP example, we notice that *lesion* is tagged in one example as *tumour* and in others as *O*, despite the examples being very similar. In the BE example, *Moderate* is not predicted to be part of the *atrophy* finding. However, it is also not annotated as such in all similar training examples below, thus highlighting how some errors can be explained by identifying inconsistencies in annotations.

For such cases where the training set contains many alternative possible labellings of tokens in particular contexts, we propose visualising the uncertainty by plotting the entropy of the kNN distribution along the sequence together with the subset of labels deemed plausible from the retrieved training examples. Figure 5 demonstrates how the boundaries of the *small vessel disease* finding are uncertain in the training set, with some instances including *periventricular* as part of the entity, and others tagging *white* as *O* in similar contexts.

To conclude, BlueBERT's pretrained preconceptions about which contexts are similar makes it harder for the model to fit examples that are annotated inconsistently with respect to spans or labels. We believe it therefore to be an effective model for fine-grained error analysis as well as for assisting in annotation efforts in tandem with any rule-based or other developed system when generating annotations in a new domain.

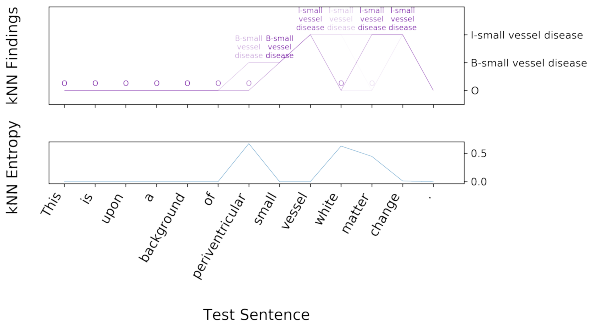


Figure 5: Distribution over *findings* $p_{kNN}$ computed using $k$=10 most similar training examples to highlight uncertainty in conflicting annotations. The top subfigure demonstrates plausible labels, with solid lines linking more likely candidates. The bottom subfigure is a plot of the entropy of $p_{kNN}$, with higher entropy corresponding to choices that are more uncertain.

### 6.3 Spelling Errors

Spelling variation and spelling mistakes are not uncommon in radiology reports. For example, the ESS data contains frequent mentions of the British English spelling variant *haemorrhage* but also several mentions of *hemorrhage*, its US En-

[3]Not finetuned on radiology data as part of EdIE-BERT.

[4]The nearest neighbour search is among tokens such as *lesion*, but we visualise the whole sentence for context.

glish version. It also includes spelling errors such as (*haemohhage*, *haemorrghage* and *heamrrhage*).

Known spelling variants can be handled with a few rules in a rule-based NLP system given a specific domain as brain imaging reports. However, terms containing spelling errors are unpredictable and hence more difficult to recognise using heuristics. Even though EdIE-R slightly outperforms the neural system overall, one main strength of EdIE-BiLSTM and EdIE-BERT is that they are robust towards spelling errors, since their model is context and subword structure-aware.

EdIE-R does not currently contain a separate spelling correction step but encodes a limited number of rules to deal with spelling errors and variations frequently observed in the data used for its development. As a result, it regards most words containing spelling errors as being out-of-vocabulary.

To examine how the neural systems dealt with actual spelling errors in radiology reports, we identified those appearing in gold findings and modifier annotations in the ESS data and found 24 unique annotations containing spelling errors (see Appendix B.1). 10 of them occur in the ESS validation and test data not used for training. Both EdIE-BiLSTM and EdIE-BERT were able to correctly recognise 6/10 and 5/10 annotations, respectively. When presented with the correctly spelled variants in the same context, they were able to identify 8/10 and 7/10 annotations accurately. While these examples are too few for a quantitative analysis of the robustness of both models towards spelling errors, it is clear that they can detect some of them accurately.

## 7 Summary and Conclusions

Access to annotated clinical text is a bottleneck to progress in clinical IE. While it is vital to strive for high quality gold datasets that are annotated from scratch with clear annotation guidelines, the reality of the situation is that many teams face data accessibility issues, strict time constraints and limited access to expert annotators, whose time is extremely valuable. Given finite resources, it is therefore common to leverage output from previously developed systems to speed up annotation. Through extensive error analysis, we exposed artefacts of annotations originating from experts correcting system output and recommend exploiting context-aware embedding models, such as BERT, to improve recall and ameliorate annotation inconsistencies. We are not suggesting that standards of the annotation procedure should be overlooked, but we highlight that our approach may be of value for many teams that are not in a position to label a dataset from scratch: semi-automated expert data is extremely useful under low resource settings, and therefore having a way to guide such annotation processes is valuable.

We also highlighted the pitfall of blindly trusting well-established metrics, both for ranking systems on subtasks that do not directly match the downstream task and, more importantly, in the case of generalisation, where metrics on in-sample data were misleading as to how well our neural models were capturing negation. We concur with the findings in Wu et al. (2014), negation detection is straightforward to optimise for an in-domain sample of data, but generalisation to other datasets without any adaptation is still challenging. Therefore, negation detection models should be tested across multiple datasets for generalisation.

To conclude, our rule-based system outperforms our neural network models on the limited sized in-sample dataset and generalises to an unseen dataset of radiology reports. Through a manual error analysis, we found that a large proportion of errors of our systems are due to ambiguities in annotation. Given the fairly high performance of our models, we extrapolate that we have likely distilled most of the information available in our limited labelled dataset. In future work we plan to extend our annotations to a larger dataset to further assess generalisation.

## Acknowledgements

We thank Arlene Casey and the anonymous reviewers for their comprehensive feedback and Laura Perez-Beltrachini for helpful discussions on an early version of the paper. We also wish to thank Prof. Cathy Sudlow for making the ESS data available for this research and the members of the Edinburgh Clinical NLP group[5] for their support.

This research was supported by the MRC Mental Health Data Pathfinder Award (MRC - MCPC17209). Moreover, Alex and Grover have been supported by the Alan Turing Institute via Turing Fellowships (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).

[5] https://www.ed.ac.uk/usher/clinical-natural-language-processing

## References


Beatrice Alex, Claire Grover, Richard Tobin, Cathie Sudlow, Grant Mair, and William Whiteley. 2019. Text mining brain imaging reports. *Journal of Biomedical Semantics*, 10(1):23.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2019. Joint entity extraction and assertion detection for clinical text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 954–959, Florence, Italy. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Wendy Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 17–27, Auxtin, TX. Association for Computational Linguistics.

James Curran and Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada*, pages 164–167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Jenny Rose Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, 6(S-1).

Sunyang Fu, Lester Y. Leung, Yanshan Wang, Anne-Olivia Raulli, David F. Kallmes, Kristin A. Kinsman, Kristoff B. Nelson, Michael S. Clark, Patrick H. Luetmer, Paul R. Kingsbury, et al. 2019. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Medical Informatics*, 7(2):e12109.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1019–1027.

Philip John Gorinski, Honghan Wu, Claire Grover, Richard Tobin, Conn Talbot, Heather Whalley, Cathie Sudlow, William Whiteley, and Beatrice Alex. 2019. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches. *Computing Research Repository*, arXiv:1903.03985. Version 2.

Claire Grover and Richard Tobin. 2006. Rule-based chunking and reusability. In *Proceedings of LREC 2006*, pages 873–878.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Saeed Hassanpour and Curtis P. Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66:29–39.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.

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. 2016.

MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*.

Chulho Kim, Vivienne Zhu, Jihad Obeid, and Leslie Lenert. 2019. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLOS ONE*, 14(2):1–13.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182.

Yoon Kim, Yacine Jernite, David A. Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2741–2749. AAAI Press.

Curtis P. Langlotz. 2006. Radlex: a new method for indexing online educational materials. *Radiographics*, 26(6):1595–1597.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020. Does BERT need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.

Christopher Manning. 2006. Doing Named Entity Recognition? Don't optimize for F1. https://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html[Accessed: 29/01/2020].

Timothy Miller, Steven Bethard, Hadi Amiri, and Guergana Savova. 2017. Unsupervised domain adaptation for clinical negation detection. In *BioNLP 2017*, pages 165–170, Vancouver, Canada,. Association for Computational Linguistics.

Guido Minnen, John Carroll, and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of INLG 2000*, pages 201–208.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Nidhin Nandhakumar, Ehsan Sherkat, Evangelos E. Milios, Hong Gu, and Michael Butler. 2017. Clinically significant information extraction from radiology reports. In *Proceedings of the 2017 ACM Symposium on Document Engineering*, DocEng '17, page 153–162, New York, NY, USA. Association for Computing Machinery.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:188—196.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. 2016. Natural Language Processing in Radiology: A Systematic Review. *Radiology*, 279(2):329–343.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada*, pages 142–147.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. 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.

Patrick Schrempf, Hannah Watson, Shadia Mikhael, Maciej Pajak, Matúš Falis, Aneta Lisowska, Keith W. Muir, David Harris-Birtill, and Alison Q.

O'Neil. 2020. Paying per-label attention for multi-label extraction from radiology reports. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 277–289, Cham. Springer International Publishing.

Thomas Searle, Zeljko Kraljevic, Rebecca Bendayan, Daniel Bean, and Richard Dobson. 2019. MedCAT-Trainer: A biomedical free text annotation interface with active learning and research use case specific customisation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 139–144, Hong Kong, China. Association for Computational Linguistics.

Elena Sergeeva, Henghui Zhu, Amir Tahmasebi, and Peter Szolovits. 2019. Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187, Hong Kong. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *Computing Research Repository*, arXiv:2004.09167. Version 2.

Dominic Sykes, Andreas Grivas, Claire Grover, Richard Tobin, Cathie Sudlow, William Whiteley, Andrew MacIntosh, Heather Whalley, and Beatrice Alex. 2020. Comparison of Rule-Based and Neural Network Models for Negation Detection in Radiology Reports. *Journal of Natural Language Engineering*. Accepted for publication.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):1–9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Computing Research Repository*, arXiv:1910.03771. Version 4.

David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, and Thomas C. Booth. 2020. Automated labelling using an attention model for radiology reports of MRI scans (ALARM). volume 121 of *Proceedings of Machine Learning Research*, pages 811–826, Montreal, QC, Canada. PMLR.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: Generalizability versus optimizability in clinical natural language processing. *PLOS ONE*, 9(11):1–11.

Henghui Zhu, Ioannis Ch. Paschalidis, Christopher Hall, and Amir Tahmasebi. 2019. Context-driven concept annotation in radiology reports: Anatomical phrase labeling. *AMIA Summits on Translational Science Proceedings*, 2019:232.

# A Supplemental Material

## A.1 Training and hyperparameter details

### A.1.1 EdIE-BiLSTM

As mentioned in Section 3.2, we employ a character level convolutional network following a modified version of the small CNN encoder model of Kim et al. (2016), details of which can be seen tabulated in Table 6. We replace tanh non-linearities with ReLUs. We also remove the highway layer since we did not observe any improvements when using it. To speed up training, we apply padding-aware batch normalisation (Ioffe and Szegedy, 2015) to the convolution activations before the ReLU non-linearity.[6] We project the character aware token embedding $c_i$ to a vector of dimensionality 128 using an affine layer $\boldsymbol{P}$. Word embeddings are also of dimensionality 128. Both word and character embeddings are randomly initialised.

Following Gal and Ghahramani (2016), we randomly drop out word types with 0.5 probability for words. We also follow this approach for characters, but with a lower dropout rate of 0.1.

Optimisation-wise, we trained our model using stochastic gradient descent with a batch size of 16 sentences padded to maximum length, a learning rate of 1 and a linear warmup of the learning rate over the first 200 parameter updates = 1 checkpoint. Before performing backpropagation, we clip the norm of the global gradient of the parameters to 5. We stop training when entity prediction does not improve on the validation set for 10 consecutive checkpoints. Our model is implemented using PyTorch (Paszke et al., 2019).

### A.1.2 EdIE-BERT

For the BlueBERT encoder we use the uncased base model trained on PubMed and MIMIC-III. We train the model using the Adam optimiser with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 16. We follow the original BERT paper and train using a warmup linear schedule, increasing the learning rate linearly for the first 400 training steps (10% of training steps) until it reaches the maximum value ($5 \cdot 10^{-5}$) and then decreasing it for the remaining 90% of training steps. A step is a parameter update, namely a forward and backward propagation of a batch. 200 parameter updates roughly correspond to 15 epochs on our ESS training set. We chose the aforementioned hyperparameter values by conducting a search over learning rate $\{5 \cdot 10^{-5}, 2 \cdot 10^{-5}\}$, batch size $\{16, 32\}$ and number of warmup steps $\{200, 400\}$ on the development set. We use the Huggingface (Wolf et al., 2020) implementation for BERT.

| Hyperparameter | Value |
|---|---|
| Char embedding dim | 15 |
| Char CNN filter widths | [1, 2, 3, 4, 5, 6] |
| Char CNN number filters | [25, 50, 75, 100, 125, 150] |
| Char project dim | 128 |
| Word embedding dim | 128 |
| BiLSTM dim | 512 |
| MLP hidden dim | 512 |
| MLP dropout | 0.25 |
| Word type dropout | 0.5 |
| Char type dropout | 0.1 |
| Non-linearity | ReLU |
| Optimiser | SGD |
| Learning rate | 1 |
| Batch size | 16 |
| Gradient clipping global norm | 5 |

Table 6: EdIE-BiLSTM hyperparameter choice.

# B Spelling Errors

## B.1 List of ESS Data Annotations with Spelling Errors

*basal galnglia*, *basal ganglia*, *centrum semiovale*, *Esatblished*, *exta-axial collections*, *extra-axia collection*, *extraxial collection*, *haemorraghic transformation*, *infarcion*, *Low attenuation of perventricular white matter*, *microvacular ischaemia*, *mircovascular ischaemia*, *parietooccpital*, *perfusion defecit*, *periventicular low attenuation*, *periventricualr white matter hypoattenuation*, *posterior cerberal artery*, *resticted diffusion*, *thebasal ganglia*, *craniopharyngoma*, *lacumar*, *brainstsem*, *occiptal* and *subdural haemohhage*

## B.2 Synthetic Spelling Error Analysis

The example in Figure 6 shows EdIE-Viz output of EdIE-BiLSTM for a synthetic report with a number of deliberately inserted spelling errors.[7]

The report contains misspellings due to character and whitespace insertions (*vesssel*, *heamorrrhage*, *a cute*), character deletions (*hypoattenuation*, *infaret*, *atrophy*, *disease*, *infarcts*, *stroke*) or character substitutions (e→a: *pariatal*, ae→ea: *heamorrrhage*). EdIE-BiLSTM is able to recognise most of the misspelled entities, with the exception of *atrophy*. As expected, EdIE-R was only able

[6] We adapt the mean and variance computation of each batch to only consider tokens that do not consist of padding.

[7] EdIE-Viz is a web-based interface to our IE models (see Appendix C).

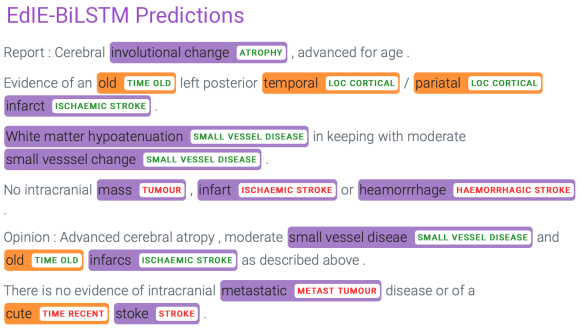


Figure 6: EdIE-BiLSTM output for a synthetic brain imaging report containing a series of spelling errors.

to tag the term *stroke* based on one of its rules allowing for that error to occur. In the case of the white-space insertion splitting *acute* into two valid English words *a cute*, EdIE-BiLSTM interestingly tags the word *cute* correctly as the temporal modifier *recent*, even though the span is wrong. Such a neural system may therefore wrongly tag a word similar in spelling but different in meaning to a medical term it is trained to extract. EdIE-BERT is able to recognise most of the misspelled findings and modifiers in this report and only differs in three cases to EdIE-BiLSTM. It is able to identify *atropy* as atrophy, does not recognise *White matter hypoatenuation* as small vessel disease and does not mark up *cute* as a modifier, presumably because during pretraining it has picked up that *cute* is a word that occurs in a different context.

## C EdIE-Viz: Interactive web demo

Our interactive web demo provides a user interface to all three systems.

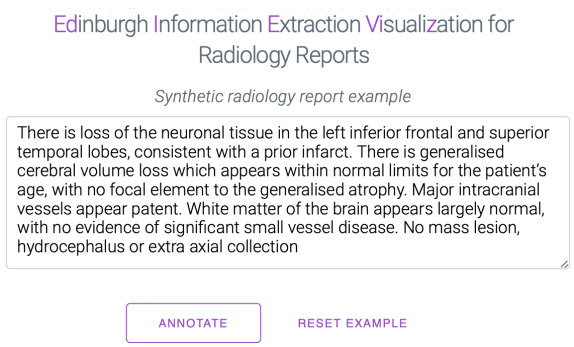


Figure 7: Home screen.

Figure 7 shows the home screen with a preloaded synthetic example of a brain imaging report. By clicking on the "Annotate" button, the demo displays[8] predicted findings (spans highlighted in purple and types displayed behind each span in all-caps), modifiers (highlighted in orange and types in all-caps) and negation (red types for negated annotations and green types for non-negated annotations) (see Figure 8). In this example, EdIE-BERT misses the negation of *small vessel disease*.

[8]The visualisation follows the style of the displaCy Named Entity Visualiser. spacy.io/usage/visualizers


Figure 8: Predicted findings, modifiers and negation for EdIE-R and EdIE-BERT. EdIE-BiLSTM output is omitted; it is identical to that of EdIE-R for this example.

We differentiate between findings and modifiers as they are notionally different (each modifier can be mapped to a finding) and because some tokens are tagged as both. For example, the abbreviation *POCI* (posterior circulation infarct) is tagged as *ischaemic stroke* and *cortical*.

The current use cases of this interface are the research team's own error analysis and system development, visual output analysis by example and system demonstrations to collaborators. However, in future it could be modified to allow bespoke processing of brain imaging reports, for example for assisting radiologists, or extended to add functionality that allows the comparison of other systems doing similar processing.

## D Availability of Data

The annotated ESS data has much potential value as a resource for developing text mining algorithms. This data will be available on application to Prof. Cathie Sudlow (email: Cathie.Sudlow AT ed.ac.uk) to bona fide researchers with a clear analysis plan.