Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks

Savelie Cornegruta, Robert Bakewell, Samuel Withey and **Giovanni Montana**
Department of Biomedical Engineering, King's College London, UK
giovanni.montana@kcl.ac.uk

Abstract

Motivated by the need to automate medical information extraction from free-text radiological reports, we present a bi-directional long short-term memory (BiLSTM) neural network architecture for modelling radiological language. The model has been used to address two NLP tasks: medical named-entity recognition (NER) and negation detection. We investigate whether learning several types of word embeddings improves BiLSTM's performance on those tasks. Using a large dataset of chest x-ray reports, we compare the proposed model to a baseline dictionary-based NER system and a negation detection system that leverages the hand-crafted rules of the NegEx algorithm and the grammatical relations obtained from the Stanford Dependency Parser. Compared to these more traditional rule-based systems, we argue that BiLSTM offers a strong alternative for both our tasks.

1 Introduction

Radiological reports represent a large part of all Electronic Medical Records (EMRs) held by medical institutions. For instance, in England alone, upwards of 22 million plain radiographs were reported over the 12-month period from March 2015 (NHS, 2016). A radiological report is a written document produced by a Radiologist, a physician that specialises in interpreting medical images. A report typically states any technical factors relevant to the acquired image as well as the presence or absence of radiological abnormalities. When an abnormality is noted, the Radiologist often gives further description, including anatomical location and the extent of the disease.

Whilst Radiologists are taught to review radiographs in a systematic and comprehensive manner, their reporting style can vary quite dramatically (Reiner and Siegel, 2006) and the same findings can often be described in a multitude of different ways (Sobel et al., 1996). The radiological reports may contain broken grammar and misspellings, which are often the result of voice recognition software or the dictation-transcript method (McGurk et al., 2014). Applying text mining techniques to these reports poses a number of challenges due to extensive variability in language, ambiguity and uncertainty, which are typical problems for natural language.

In this work we are motivated by the need to automatically extract standardised clinical information from digitised radiological reports. A system for the fully-automated extraction of this information could be used, for instance, to characterise the patient population and help health professionals improve day-to-day services. The extracted structured data could also be used to build management dashboards (Simpao et al., 2014) summarising and presenting the most prevalent conditions. Another potential use is the automatic labelling of medical images, e.g. to support the development of computer-aided diagnosis software (Shin et al., 2015).

In this paper we propose a recurrent neural network (RNN) architecture for modelling radiological language and investigate its potential advantages on two different tasks: medical named-entity recognition (NER) and negation detection. The model, a bi-directional long short-term memory (BiLSTM) network, does not use any hand-engineered features,

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 17–27,
Austin, TX, November 5, 2016.

but learns them using a relatively small amount of labelled data and a larger but unlabelled corpus of radiological reports. In addition, we explore the combined use of BiLSTM with other language models such as GloVe (Pennington et al., 2014) and a novel variant of GloVe, proposed here, that makes use of a medical ontology. The performance of the BiLSTM model is assessed comparatively to a rule-based system that has been optimised for the tasks at hand and builds upon well established techniques for medical NER and negation detection. In particular, for NER, the system uses a baseline dictionary-based text mining component relying on a curated dictionary of medical terms. As a baseline for the negation detection task, the system implements a hybrid component based on the NegEx algorithm (Chapman et al., 2013) in conjunction with grammatical relations obtained from the Stanford Dependency Parser (Chen and Manning, 2014).

The article is organised as follows. In Section 2 we provide a brief review of the existing body of work in NLP for medical information extraction and briefly discuss the use of artificial neural networks for NLP tasks. In Section 3 we describe the datasets used for our experiments, and in Section 4 we introduce the BiLSTM model. The results are presented in Section 6 where we also compare BiLSTM against the rule-based baseline systems described in Section 5.

2 Related Work

2.1 Medical NER

A large proportion of NLP systems for medical text mining use dictionary-based methods for extracting medical concepts from clinical document (Friedman et al., 1995; Johnson et al., 1997; Aronson, 2001; Savova et al., 2010). The dictionaries that contain the correspondence between a single- or multi-word phrase and a medical concept are usually built from medical ontologies such as the Unified Medical Language System (UMLS) (NLM, 2016b) and Medical Subject Headings (MeSH) (NLM, 2016a). These ontologies contain hundreds of thousands of medical concepts. There are also domain-specific ontologies such as RadLex (Langlotz, 2006), which has been developed for the Radiology domain, and currently contains over $68,000$ concepts.

Medical Language Extraction and Encoding System (MEDLEE) (Friedman et al., 1995) is one of the earliest automated systems originally developed for handling radiological reports, and later expanded to other medical domains. MEDLEE parses the given clinical documents by string matching: the words are matched to a pre-defined dictionary of medical terms or semantic groups (e.g. *Central Finding*, *Bodyloc Modifier*, *Certainty Modifier* and *Region Modifier*). Once the words have been associated with a semantic group, a Compositional Regularizer stage combines them according to a list of pre-defined mappings to form regularized multi-word phrases. The final stage looks up the regularized terms in a dictionary of medical concepts (e.g. *enlarged heart* is mapped to the corresponding concept *cardiomegaly*). A separate study evaluated MEDLEE on 150 manually annotated radiology reports (Hripcsak et al., 2002); MEDLEE was assessed on its ability to detect 24 clinical conditions achieving an average sensitivity and specificity of 0.81 and 0.99, respectively.

A more recent system for general medical information extraction is the Mayo Clinic's Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010), which also implements an NLP pipeline. During an initial shallow parsing stage, cTAKES attempts to group words into multi-word expressions by identifying constituent parts of the sentence (e.g. noun, prepositional, and verb phrases). It then string matches the identified phrases to a concept in UMLS. A new set of semantic groups were also derived from the UMLS ontology (Ogren et al., 2007). The NER performance of the cTAKES was evaluated on the semantic groups, achieving an F1-score of 0.715 for exact matches and 0.824 for overlapping matches.

In general, dictionary-based systems perform with high precision on the NER tasks but have a low recall, showing a lack of generalisation. Low recall is usually caused by the inability to identify multi-word phrases as concepts, unless exact matches can be found in the dictionary. In addition, such systems are not able to easily deal with disjoint entities. For instance, in the phrase *lungs are mildly hyperexpanded*, *hyperexpanded lungs* constitutes a clinical finding. In an attempt to deal with disjoint entities, rule-based systems such as MEDLEE,

MetaMap (Aronson, 2001) and cTAKES, implement additional parsing stages to find grammatical relations between different words in a sentence, thus aiming to create disjoint multi-word phrases. However, state-of-the-art syntactic parsers are still likely to fail when parsing sentences with broken grammar, as often occurs in clinical documents.

In an attempt to improve upon dictionary-based information extraction systems, Hassanpour (2015) recently used a first-order linear-chain Conditional Random Field (CRF) model (Lafferty et al., 2001) in a medical NER task involving five semantic groups (anatomy, anatomy modifier, observation, observation modifier, and uncertainty). The features used for the CRF model included part-of-speech (POS) tags, word stems, word n-grams, word shape, and negations extracted using the NegEx algorithm. The model was trained and tested using 10-fold cross validation on a corpus of 150 multi-institutional Radiology reports and achieved a precision score of 0.87, recall of 0.84, and F1-score of 0.85.

2.2 Medical negation detection

NegEx, a popular negation detection algorithm, is usually applied to medical concepts after the entity recognition stage. This tool uses a curated list of phrases (e.g. *no*, *no sign of*, *free of*), which are string matched to the medical text to detect a negation trigger, i.e. a word or phrase indicating the presence of a negated medical entity in the sentence. The target entities falling inside a window, starting at the negation trigger, are then classified as *negated*. In light of its simplicity, speed and reasonable results, NegEx had been used as a component by many medical NLP systems (Wu et al., 2014). It has been shown that that NegEx achieves an accuracy of 0.94 as part of the cTAKES evaluation (Savova et al., 2010). However, the window approach that is used for classifying the negations may result in a large number of false positives, especially if there are multiple entities within the 6-word window.

Aiming to reduce the number of false positives, recent efforts have integrated NegEx with machine learning models that can be trained on annotated datasets. For instance, Shivade (2015) introduced a kernel-based approach that uses features built using the type of negation trigger, features that are derived from the existence of conjunctions in the sentence, and features that weight the NegEx output against the bag-of-words in the dataset. The kernel based model outperformed the original NegEx algorithm by 2.7 F1-score points when trained and tested on the NegEx dataset. At around the same time, Mehrabi (2015) introduced DEEPEN, an algorithm that filters the NegEx output using the grammatical relations extracted using Stanford Dependency Parser. DEEPEN succeeded at reducing the number of false positives, although it showed a marginally lower F1-score when compared with NegEx on concepts from the *Disorders* semantic group from the Mayo Clinic dataset (Ogren et al., 2007).

2.3 Neural networks for NLP tasks

In recent years, deep artificial neural networks have been found to yield consistently good results on various NLP tasks. The SENNA system (Collobert et al., 2011), which used a convolutional neural network (CNN) architecture, came close to achieving state-of-the-art performance across the tasks of POS tagging, shallow parsing, NER, and semantic role labeling. More recently, recurrent neural networks (RNNs) have been shown to achieve very high performance, and often reach state-of-the-art results in various language modelling tasks (Mikolov and Zweig, 2012). RNNs have also been shown to outperform more traditional machine learning models, such as Logistic Regression and CRF, at the slot filling task in spoken language understanding (Mesnil et al., 2013). In a NER task on the publicly available datasets in four languages, the bidirectional long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), a variant of RNN, outperformed CNNs, CRFs and other models (Lample et al., 2016).

Neural networks have also been used to learn language models in an unsupervised learning setting. Some popular models include Skip-gram and continuous bag-of-words (CBOW) (Mikolov et al., 2013). These yield word representations, or embeddings, that are able to carry the syntactic and semantic information of a language. Collobert (2011) showed that integrating pre-trained word embeddings into a neural network can help the supervised learning process.

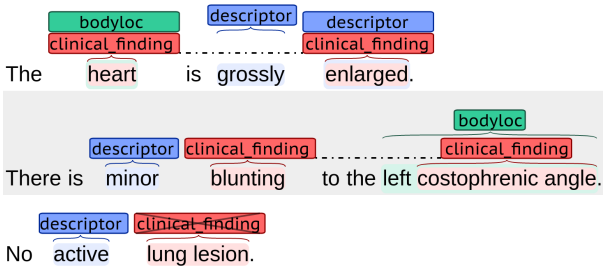


Figure 1: Example of manual annotation of a radiology report performed using BRAT

Semantic Group	# of entities	# of tokens
Body Location	5686	10113
Clinical Finding	5396	8906
Descriptor	3458	3845
Medical Device	1711	3361
Total	16251	26225
Negated entities	1851	2557

Table 1: Frequency distribution of entities by class in $2,000$ manually annotated reports

3 A Radiology corpus

3.1 Dataset

For this study, we produced an in-house radiology corpus consisting of $745,480$ historical chest X-ray (radiographs) reports provided by Guy's and St Thomas' Trust (GSTT). This Trust runs two hospitals within the National Health Service (NHS) in England, serving a large area in South London. The reports cover the period between January 2005 and March 2016, and were generated by 276 different reporters including consultant Radiologists, trainee Radiologists and reporting Radiographers. Our repository consists of text written or dictated by the clinicians after radiograph analysis, and do not contain any referral information or patient-identifying data, such as names, addresses or dates of birth. However, many reports refer to the clinical history of the patient. The reports had a minimum of 1 word and maximum of 311 words, with an average of 25.3 words and a standard deviation of 19.9 words. On average there were 2.9 sentences per report. After lemmatization, converting to lower case, and discounting words that occur less than 3 times in the corpus, the resulting vocabulary contained $8,031$ words.

A sample of $2,000$ reports was randomly selected from the corpus for the purpose of creating a training and validation dataset for the NER and negation detection tasks, whilst the remaining of the reports were utilised for pre-training word embeddings. The reports selected for manual annotation were written for all types of patients (Inpatient: 1072, A&E Attender: 515, Outpatient: 229, GP Direct Access Patient: 165, Ward Attender: 9, Day Case Patient: 8) by 144 different clinicians.

We introduce a simple word-level annotation schema that includes four classes or semantic groups: *Clinical Finding*, *Body Location*, *Descriptor* and *Medical Device*: Clinical Finding encompasses any clinically-relevant radiological abnormality, Body Location refers to the anatomical area where the finding is present, and Descriptor includes all adjectives used to describe the other classes. The Medical Device class is used to label any medical apparatus seen on chest radiographs, such as pacemakers, intravascular lines, and nasogastric tubes. Our annotation schema allows for the same token to belong to several semantic groups. For example, as shown in Figure 1, the word *heart* was associated with both Clinical Finding and Body Location classes. We have also introduced a negation attribute to indicate the absence of any of these entities.

3.2 Gold standard

Two clinicians (RB and SW) annotated the reports using BRAT (Stenetorp et al., 2012), a collaborative tool for text annotation that was configured to use our own schema. The BRAT output was then transformed to the IOBES tagging schema. Here, we interpret I as a token in the middle of an entity; O as a token not part of the entity; B and E as the beginning and end of the entity, respectively; finally, S indicates a single-word entity. We work with the assumption that entities may be disjoint and tokens that are surrounded by disjoint entity may belong to a different semantic group. For example, according to the annotation performed by the clinicians, in the sentence *Heart is slightly enlarged* the phrase *heart enlarged* represents an entity that belongs to the semantic group *Clinical Finding* and *slightly* is a *Descriptor*. The resulting breakdown of all entities by semantic group can be found in Table 1.

4 Methodology

In this Section we describe a model for NER that extracts five types of entities: the four semantic groups described in Section 3.1, as well as the negation, which is treated here as an additional class, analogously to the semantic groups.

4.1 Bi-directional LSTM

The RNN is a neural network architecture designed to model time series, but it can be applied to other types of sequential data (Rumelhart et al., 1988). As the information passes through the network, it can persist indefinitely in its memory. This facilitates the process of capturing sequential dependencies. The RNN makes a prediction after processing each element of the input sequence. Hence, the output sequence can be of the same length as the input sequence. The RNN architecture lends itself as a natural model for the proposed NER task, where the objective is to predict the IOBES tags for each of the input words.

The RNN is trained using the error backpropagation through time algorithm (Werbos, 1990) and a variant of the gradient descent algorithm. However, training these models is notoriously challenging due to the problem of exploding and vanishing gradients, especially when trained with long input sequences (Bengio et al., 1994). For the exploding gradient problem, numerical stability can be achieved by clipping the gradients (Graves, 2013). The problem of vanishing gradients can be addressed by replacing the standard RNN cell with a long short-term memory (LSTM) cell, which allows for a constant error flow along the input sequence (Hochreiter and Schmidhuber, 1997). A more constant error also means that the network is able to learn better long-term dependencies over the input sequence. By combining the outputs of two RNNs that pass the information in opposing directions, it is possible to capture the context from both ends of the sequence. The resulting architecture is known as Bidirectional LSTM (BiLSTM) (Graves and Schmidhuber, 2005).

We start by defining a vocabulary $V = \{v_1, v_2, ..., v_{8031}\}$ that contains the words extracted from the corpus as described in Section 3.1. We assume that, in order to perform NER on the words in any given sentence, it is sufficient to consider only the information contained in that sentence. Therefore we pass the BiLSTM one sentence at a time. For each input sentence of n words we define an n-dimensional vector $\mathbf{x}$ whose elements are the indices in V corresponding to words appearing in the sentence, preserving the order. The input $\mathbf{x}$ is passed to an Embedding Layer that returns the sequence $S = \{w_j | j = x_1, x_2, ..., x_n\}$ where w_j is the jth row of a dense matrix $\mathbf{W} \in \mathbb{R}^{|V| \times d}$, where $d \in \mathbb{N}$ is a hyperparameter. The vector w_j represents a low-dimensional vector representation, or word embedding, whereas $\mathbf{W}$ is the corresponding embedding matrix. The sequence of word embeddings S is then passed as input to two LSTM layers that process it in opposing directions (forwards and backwards), similar to the architecture introduced by Graves (2005). Figure 2 shows the LSTM layers in their "unrolled" form as they read the input. Each LSTM layer contains k LSTM memory cells which are based on the implementation by Graves (2013). The output from each of the LSTM layers is $H = \{\mathbf{h}_t \in \mathbb{R}^k | t = 1, 2, ..., n\}$.

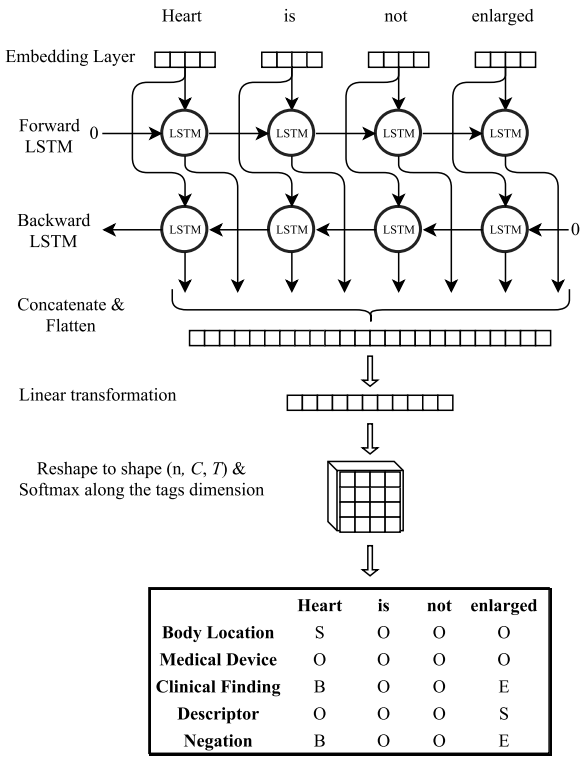
Next, we concatenate and flatten $H_{forward}$ and $H_{backward}$, obtaining a vector $\mathbf{p} \in \mathbb{R}^{2kn}$. We pass $\mathbf{p}$ through a linear transformation layer and reshape its output to a tensor of size $n \times C \times T$, where C is the number of annotation classes (5 in total, 4 semantic groups and 1 class for negation) and T is the number of possible tags (5 for the IOBES tags). Finally we apply the softmax function along the last dimension of the tensor to approximate the probability for each of the possible tags for each of the annotation class.

4.2 Word embeddings

We explored 4 different techniques for learning word embeddings from the text. The embeddings will subsequently be used to initialise the embedding matrix $\mathbf{W}$ that is required by BiLSTM for the NER task. In previous work, the initialisation of $\mathbf{W}$ with pre-trained embeddings has been found to improve the training process (Collobert et al., 2011; Mesnil et al., 2013).

Random Embeddings

Random embeddings were obtained by drawing from a uniform distribution in the $(-0.01,\ 0.01)$ range. As such, the positions of the words in the vector space do not provide any information regard-



	Heart	is	not	enlarged
Body Location	S	O	O	O
Medical Device	O	O	O	O
Clinical Finding	B	O	O	E
Descriptor	O	O	O	S
Negation	B	O	O	E

Figure 2: An illustration of the BiLSTM architecture for joint medical entity recognition and negation detection

ing patterns of relationships between words.

BiLSTM Embeddings

These embeddings were obtained after adapting the BiLSTM for a language modelling task. Following a previously described strategy (Collobert and Weston, 2008), the input words were randomly replaced, with probability 0.2, with a word extracted from V. We then created a corresponding vector of binary labels to be used as prediction targets: each element of the vector is either 0 or 1, where 0 indicates a word that has been replaced, and 1 indicates an unchanged word. The model outputs the probability of the labels for each word in the given sentence. After training this language model on the unlabelled part of our corpus, we extracted the word embeddings from $\mathbf{W}$.

GloVe Embeddings

Word embedding were also obtained using GloVe, an unsupervised method (Pennington et al., 2014). On word similarity and analogy tasks, it has the potential to outperform competing models such as Skip-gram and CBOW. The GloVe objective function is

$$\sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T \tilde{w} + b_i + \tilde{b}_j - logX_{ij})^2$$

where X is the word-word co-occurrence matrix, f is a weighting function, w are word embeddings, and $\tilde{w} \in \mathbb{R}^d$ are context word embeddings, with b and $\tilde{b}$ the respective bias terms. The GloVe embeddings w are trained using AdaGrad optimisation algorithm (Duchi et al., 2011), stochastically sampling nonzero elements from X.

GloVe-Ontology Embeddings

Furthermore, we introduced a modified version of GloVe, denoted GloVe-Ontology, with the objective to leverage the RadLex ontology during the word embedding estimation process. The rationale is to impose some constrains on the estimated distance between words using semantic relationships extracted from RadLex; this is an idea somewhat inspired by previous work (Yu and Dredze, 2014).

The RadLex data was initially represented as a tree, τ, by considering only the relation *is_parent_of* between concepts. We then attempted to string match every word v in V to a concept in τ. Every v matched with a RadLex concept was then assigned the vector that enumerates all ancestors of that concept; otherwise it was associated with a zero vector. We denote the resulting vector by ϕ. We imposed the constraint that words close to each other in τ should also be close in the learned embedding space. Accordingly, GloVe's original objective function was modified to incorporate this additional penalty:

$$\sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T \tilde{w} + b_i + \tilde{b}_j - logX_{ij} - \alpha sim(\phi_i, \phi_j))^2$$

In this expression, α is a parameter controlling the influence of this additional constraint, and sim is taken to be the cosine similarity function. No major changes in the training algorithm were required compared to the original GloVe methodology.

4.3 BiLSTM implementation and training

The BiLSTM was implemented using two open-source libraries, *Theano* (Theano Development

Team, 2016) and *Lasagne* (Dieleman et al., 2015). The number of memory cells in each LSTM layer, k, was set to 100. We limited the maximum length of the input sequence to 40 words and for shorter inputs we used a binary mask at the input and cropped the output predictions accordingly. The loss function was the categorical cross-entropy between the predicted probabilities of the IOBES tags and the true tags. BiLSTM was trained on a GPU for 20 epochs in batches of 10 sentences using Stochastic Gradient Descent (SGD) with Nesterov momentum and with the learning rate set to 0.5.

The embedding size d was set to 50. The GloVe, GloVe-Ontology and BiLSTM word embeddings were trained on $743,480$ unlabelled radiology reports. The α paramenter in the Glove-Ontology objective was set to 0.5.

One aspect of the training was to allow or block the optimisation algorithm from updating the matrix $\mathbf{W}$ in the Embedding Layer of the BiLSTM. In Section 6 we refer to this aspect of training as *fine-tuning*. Previous work (Collobert et al., 2011) has shown that fine-tuning can boost the results of the several supervised tasks in NLP.

5 A competing rule-based system

Two clinicians (RB and SW) built a comprehensive dictionary of medical terms. In the dictionary, the key is the name of the term and the corresponding value specifies the semantic group, which was identified using a number of resources. We iterated over all RadLex concepts using the field *Preferred Label* as the dictionary key for the new entry. To obtain the semantic group we traversed up the ontology tree until an ancestor concept was found that had been manually mapped to a semantic group. For example, one of the ancestor concepts of *heart* is *Anatomical entity*, which we had manually mapped to semantic group *Body Location*. The same procedure was also performed on the MeSH ontology using the *MeSH Heading* field as a dictionary key. Finally, we added 202 more terms that were common in day-to-day reporting but were not present in RadLex and MeSH.

The sentences were tokenized and split using the Stanford CoreNLP suite (Manning et al., 2014), and also converted to lower case and lemmatized using NLTK (Bird et al., 2009). Next, for each sentence, the algorithm attempted to match the longest possible sequence of words, a target phrase, to an entry in the dictionary of medical terms. When the match was successful, the target phrase was annotated with the corresponding semantic group. When no match was found, the algorithm attempted to look up the target phrase in the English Wikipedia redirects database. In case of a match, the name of the target Wikipedia article was checked against our curated dictionary and the target phrase was annotated with the corresponding semantic group (e.g. *oedema* redirects to *edema*, which is how this concept is named in RadLex).

For all the string matching operations we used SimString (Okazaki and Tsujii, 2010), a fast and efficient approximate string matching tool. We arbitrarily chose the *cosine* similarity measure and a similarity threshold value of 0.85. Using SimString allowed the system to match misspelled words (e.g. *cardiomegally* to the correct concept *cardiomegaly*).

For negation detection, the system first obtained NegEx predictions for the entities extracted in the NER task. Next, it generated a graph of grammatical relations as defined by the Universal Dependencies (De Marneffe et al., 2014) from the Stanford Dependency Parser. It then removed all relations in the graph except *neg*, the negation relation, and *conj:or*, the *or* disjunction. Given the NegEx output and the reduced dependency graph, the system finally classified an entity as negated if any of the following two conditions were found to be true: (1) any of the words that are part of the entity were in a *neg* relation or in a *conj:or* relation with a another word that was in a *neg* relation; (2) if an entity was classified by NegEx as negated, it was the closest entity to negation trigger and there was no *neg* relations in the sentence. Our hybrid approach is somewhat similar to DEEPEN with the difference that the latter considers all first-order dependency relations between the negation trigger and the target entity.

6 Experimental Results

We evaluated the BiLSTM model on the medical NER task by measuring the overlap between the predicted semantic groups and the ground truth labels. The evaluation was performed at the granularity of a single word and using 5-fold cross-validation. The BiLSTM model was always trained on 80% of the annotated corpus and tested on the remaining 20%.

Embeddings	Fine-tuning	P	R	F1
Random	TRUE	0.878	0.869	0.873
Glove	TRUE	0.869	0.829	0.849
Glove-ontology	TRUE	0.875	0.860	0.867
BiLSTM	TRUE	0.878	0.870	**0.874**
Random	FALSE	0.829	0.727	0.775
Glove	FALSE	0.866	0.828	0.847
Glove-ontology	FALSE	0.850	0.839	0.844
BiLSTM	FALSE	0.870	0.849	**0.859**
Rule-based		0.706	0.698	0.702

Table 2: Comparison of the BiLSTM model and rule-based system. BiLSTM is trained using different word embedding and evaluated using 5-fold cross-validation. The evaluation considers the overlap span of the semantic group predictions against gold standard annotations.

Semantic Group	P	R	F1
Body Location	0.896	0.887	0.891
Medical Device	0.898	0.923	0.910
Clinical Finding	0.871	0.895	0.883
Descriptor	0.824	0.725	0.771
Total	0.878	0.870	0.874

Table 3: BiLSTM: performance metrics broken down by semantic group for the NER task. All results were obtained using BiLSTM word embeddings.

Semantic Group	P	R	F1
Body Location	0.724	0.839	0.778
Medical Device	0.976	0.538	0.694
Clinical Finding	0.862	0.551	0.672
Descriptor	0.467	0.780	0.584
Total	0.706	0.698	0.702

Table 4: Rule-based system: performance metrics broken down by by semantic group for the NER task.

Model	P	R	F1
BiLSTM	0.903	0.912	0.908
NegEx	0.664	0.944	0.780
NegEx - Stanford	0.944	0.912	**0.928**

Table 5: Comparison of BiLSTM, NegEx and NegEx-Stanford for negation detection. All algorithms predicted whether a given medical entity was negated or affirmed.

Table 2 compares the performance of various BiLSTM variants that were obtained with and without fine-tuning of the word embeddings to the performance of our baseline rule-based system. Without fine-tuning, the BiLSTM NER model, that was initialised with the embeddings trained in an unsupervised manner using the BiLSTM language model, achieves the best F1-score (0.859), and outperforms the next best variant by 0.012. With fine-tuning, the same BiLSTM variant improves the F1-score by a further 0.015 and outperforms the baseline rule-based system by an F1-score of 0.172. Table 3 shows its performance measure for each of the semantic groups.

nodule	pacemaker	small	remains	fracture
bulla	ppm	tiny	remain	fractures
nodules	icd	minor	appears	deformity
opacity	wires	mild	is	body
opacities	drains	dense	are	scoliosis
opacification	leads	extensive	were	abnormality

Table 6: For each one of the five words in boldface, five nearest neighbours found in the embedding space learnt by BiLSTM.

The evaluation of negation detection was measured on complete entities. If any of the words within an entity were tagged with a I, B, E or S, that entity was considered to be negated. As shown in Table 5, the BiLSTM (BiLSTM language model embeddings, fine-tuning allowed) achieved an F1-score of 0.902, which outperformed NegEx by 0.128. However, the best F1-score of 0.928 is achieved using the NegEx-Stanford system.

7 Discussion

In Table 3, we show the predictive performance of the best BiLSTM NER model for each of the semantic groups. *Body Location*, *Medical Device* and *Clinical Finding* show a balanced precision and recall, and similar F1-scores. *Descriptor* has a lower F1-score which is caused by a low recall that may be the results of the larger variability in the words used for this semantic group. Table 4 shows the corresponding results for the rule-based NER system. *Medical Device* and *Clinical Finding* show a typical performance for a dictionary-based NER system with a high precision and a low recall. *Body Location* has relatively high precision and recall values which suggests that this semantic group is well covered by our dictionary of medical terms. In contrast, *Descriptor* shows a very low precision which is the result of a high number of false positives. The false

positives are caused by many *Descriptor* entries in our dictionary of medical terms that had been automatically extracted from RadLex and MeSH but which do not correspond to the definition of a *Descriptor* used by the clinicians who produced the labelled data.

As a qualitative assessment, Table 6 shows the 5 nearest neighbours obtained from BiLSTM language model embeddings of some frequent words used by Radiologists. We note that there is an clear semantic similarity between the nearest neighbour words. Additionally, the embeddings encode syntactic information as the nearest neighbour words are parts of speech of the same type as the target word. We also summed the vectors for *heart* and *enlarged*, which yielded vec(*cardiomegaly*) as the nearest vector. Similarly, the closest vector to vec(*heart*) + vec(*not*) + vec(*enlarged*) is vec(*normal*). These examples suggest that word embeddings may encode information about the compositionality of words as discussed by Mikolov (2013).

Table 2 shows that, without fine-tuning, the Embedding Layer weights can affect the performance of the NER task. When fine-tuning is allowed there is only a marginal advantage in using pre-trained embeddings, as the BiLSTM performs equally well when initialised with random embeddings. Therefore, despite a positive qualitative assessment, the pre-trained word embeddings seem to offer only a small advantage when used for the proposed NER task as BiLSTM is able to learn well using the annotated data during the supervised learning phase.

8 Conclusions

In this paper we have shown that a recurrent neural network architecture, BiLSTM, can learn to detect clinical findings and negations using only a relatively small amount of manually labelled radiological reports. Using a manually curated medical corpus, we have provided initial evidence that BiLSTM outperforms a dictionary-based system on the NER task. For the detection of negations, on our dataset BiLSTM approaches the performance of a negation detection system that was build using the popular NegEx algorithm and uses grammatical relations obtained from the Stanford Dependency Parser and hand-crafted rules. We believe that increasing the size of the annotated training dataset can result in much improved performance on this task, and plan to purse this investigation in future work.

We have also investigated potential performance gains that can be achieved by using pre-trained word embeddings, i.e. BiLSTM, GloVe and GloVe-Ontology embeddings, in the context of BiLSTM-based modelling for the NER task. Our initial experimental results suggest that there is marginal benefit in using BiLSTM-learned embeddings while pre-training using GloVe and GloVe-Ontology embeddings did not offer any significant improvements over a random initialisation.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. " O'Reilly Media, Inc.".

Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. 2013. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics*, 192:677.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Sander Dieleman, Jan Schlter, Colin Raffel, Eben Olson, Sren Kaae Snderby, Daniel Nouri, Daniel Maturana,

Martin Thoma, Eric Battenberg, Jack Kelly, Jeffrey De Fauw, Michael Heilman, diogo149, Brian McFee, Hendrik Weideman, takacsg84, peterderivaz, Jon, instagibbs, Dr. Kashif Rasul, CongLiu, Britefury, and Jonas Degrave. 2015. Lasagne: First release., August. Available at http://dx.doi.org/10.5281/zenodo.27878 .

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(01):83–108.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Saeed Hassanpour and Curtis P Langlotz. 2015. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

George Hripcsak, John HM Austin, Philip O Alderson, and Carol Friedman. 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. *Radiology*, 224(1):157–163.

David B. Johnson, Ricky K. Taira, Alfonso F. Cardenas, and Denise R. Aberle. 1997. Extracting information from free text radiology reports. *Int. J. Digit. Libr.*, 1(3):297–308, December.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Curtis P Langlotz. 2006. Radlex: a new method for indexing online educational materials 1. *Radiographics*, 26(6):1595–1597.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

S McGurk, K Brauer, TV Macfarlane, and KA Duncan. 2014. The effect of voice recognition software on comparative error rates in radiology reports. *The British journal of radiology*.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, 54:213–219.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775.

Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *SLT*, pages 234–239.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

England NHS. 2016. Diagnostic Imaging Dataset Statistical Release. Available at https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2015-16-data/.

United States National Library of Medicine NLM. 2016a. Medical Subject Headings. Available at https://www.nlm.nih.gov/mesh/.

United States National Library of Medicine NLM. 2016b. Unified Medical Language System. Available at https://uts.nlm.nih.gov/home.html.

Philip V Ogren, Guergana K Savova, Christopher G Chute, et al. 2007. Constructing evaluation corpora for automated clinical named entity recognition. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 2325. IOS Press.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 851–859. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Bruce Reiner and Eliot Siegel. 2006. Radiology reporting: returning to our image-centric roots. *American Journal of Roentgenology*, 187(5):1151–1155.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1988. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

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.

Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. 2015. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1099.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Extending NegEx with kernel methods for negation detection in clinical text. *ExProM 2015*, page 41.

Allan F Simpao, Luis M Ahumada, Jorge A Gálvez, and Mohamed A Rehman. 2014. A review of analytics and clinical informatics in health care. *Journal of medical systems*, 38(4):1–7.

Jeffrey L Sobel, Marjorie L Pearson, Keith Gross, Katherine A Desmond, Ellen R Harrison, Lisa V Rubenstein, William H Rogers, and Katherine L Kahn. 1996. Information content and clarity of radiologists' reports for chest radiography. *Academic radiology*, 3(9):709–717.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May.

Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negations not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *ACL (2)*, pages 545–550.