

Extracting Supporting Evidence from Medical Negligence Claim Texts

Robert Bevan[†], Alessandro Torrissi[†], Danushka Bollegala[†], Frans Coenen[†], Katie Atkinson[†]

[†] University of Liverpool

{robert.bevan, alessandro.torrissi, danushka, coenen, k.m.atkinson}@liverpool.ac.uk

Abstract

The number of medical negligence claims filed in the UK each year has increased significantly over the past decade [NHS, 2018]. When filing a case of medical negligence, electronic health records act as a legally valid important source of evidence. Patients often undergo different and complex treatments over many months or years, easily resulting in hundreds of pages of electronically available medical records. Therefore, it is a non-trivial task to read all the related electronic health records and identify the supporting evidence to establish a legal case. Currently, the process of identifying evidence is carried out by humans who are experts in both medical negligence law as well as medicine. In this paper, we compare different methods of automatically extracting relevant statements from medical negligence claim texts, to move towards building a method for extracting relevant sections from electronic health records with the aim of expediting the litigation process and reducing the manual efforts involved. Specifically, we annotate a dataset containing medical negligence claim texts and train conditional random field (CRF) and long short-term memory (LSTM) network models for extracting information relevant to cases. Our evaluation shows that each model class has its merits in this task: the CRF models were significantly more effective in identifying full sequences, while the LSTMs were significantly better at assigning tags to tokens. We found both approaches were able to identify information that is key to the litigation process.

1 Introduction

Medical negligence claims are a significant source of litigation. For example, in 2018, the national health service (NHS) in the United Kingdom reported that it paid GBP 1,623 million as compensation for 10,637 claims [NHS, 2018]. Acts of medical negligence can vary in complexity as well as severity. Finding the reasons behind medical negligence acts is important in order to prevent such unfortunate events in the future [Toyabe, 2012]. Moreover, in the event where a patient

*University hospital **mistakenly amputated my left leg** despite the fact the cancer was confined within my right leg. I will now need to undergo **another leg amputation** and will be **confined to a wheelchair for the rest of my life.***

Figure 1: An example extraction (performed by a human).

(or a legal representative acting on behalf of a patient), would like to prosecute the health care provider for medical negligence, a legal case must be filed based on medical evidence. An important source of medical evidence for such prevention efforts or litigation processes is the electronic health records describing the various treatments undergone by the patient, the medication prescribed for the patient, and their medical history. The volume of electronic health records for a single patient can be significant. It is not uncommon for a patient to be subjected to medical treatment for many months, if not years, and typically a much smaller set of relevant evidence supporting the medical negligence case must be identified from this vast amount of information. A further complication is the non-contiguous distribution of evidence contained within the medical records. For example, negative patient outcomes may occur years after an initial negligent act, as a result, simply filtering the medical records according to the date of the negligent act may not be effective in building a body of evidence.

The existing process for identifying supporting evidence from electronic health records is a manual one. Humans who are knowledgeable in both medical negligence law and medicine must manually read a collection of medical records and then carefully select parts that can be used as evidence in the litigation process. Needless to say, this is both a time consuming and a costly process. Moreover, the number of individuals possessing both legal and medical background knowledge is small, which means a limited number of medical records can be read and analysed over a given period of time. These drawbacks in the existing pipeline for extracting evidence call for automatic methods that can efficiently “read” large quantities of medical records and accurately extract the relevant evidence.

In this paper, given medical negligence claim texts, we compare methods of automatically extracting expressions that are relevant to the medical negligence case: the alleged negligent acts, and any consequential negative patient outcomes.

This can be useful in helping lawyers quickly establish the key elements of the case, and we conjecture this will be useful as part of a system for automatically extracting supporting evidence from medical records.

Specifically, first we manually annotate a set of medical negligence claim texts, identifying any statements of negligent acts and any consequential negative patient outcomes. An example is shown in Figure 1, where text relating to negligent acts and negative outcomes are highlighted in red and blue respectively. Next, we train a Conditional Random Field (CRF) [Lafferty *et al.*, 2001] model for predicting BIO (Begin-Inside-Outside) tags for extracting sequences of tokens in texts belonging to the previously described categories. We use different types of features such as Part of Speech (POS), typography, and medical lexicons. One issue we encounter in this approach is the data sparseness – the limited overlap of the tokens between the training and testing data. To overcome this data sparseness issue, we use pre-trained word embeddings and automatically append training instances with related features that did not appear in the original training instances. Our experimental results show that this feature augmentation approach successfully overcomes the data sparseness problem. Finally, we train various Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997] for the same task. We experiment with both regular and Bidirectional LSTMs (BiLSTMs), and make use of both word and character level features.

2 Related Work

Information extraction has a long and established history as a task in NLP. In Named Entity Recognition (NER) [Shen *et al.*, 2018; Kuru *et al.*, 2016; Ritter *et al.*, 2011; Guo *et al.*, 2009; Rud *et al.*, 2011], the goal is to extract mentions of named entities such as people, locations, organisations, products etc. It has been reported that over 70% of web search queries contain some form of a named entity [Guo *et al.*, 2009]. Therefore, being able to recognise named entities enables us to find more relevant results in information retrieval. Relation Extraction (RE) [Mandya *et al.*, 2017; Miwa and Bansal, 2016] further extends this process by identifying the semantic relations that exist between two or more recognised named entities. For example, a *competitor* relation can exist between two companies, which can later transform into an *acquisition* relation. In medical contexts, identifying the adverse reactions associated with drugs (ADRs) from formal reporting tools, such as the Yellow Card System, or more informal reporting methods, such as social media, has received wide attention [Bollegala *et al.*, 2018; Sloane *et al.*, 2015].

Our problem: extracting litigation relevant statements from medical negligence case texts, can be seen as a specific instance of the above-described information extraction problem. However, there are some important properties in our case, which differentiate it from the more popular information extraction problems such as NER, RE or ADR extraction. First, compared to, for example, named entities, evidence related to medical negligence tends to comprise longer sequences. For example, the evidence extracted in Figure 1

Statement type	Count	Mean word count
negligent act	2551	11 (+/- 6)
negative outcome	5510	4 (+/- 3)

Table 1: Dataset summary.

Generic	Sparseness	Domain specific
word	stem	sentiment
word suffixes	stem suffixes	in_medical_lexicon
is_upper_case	similar words	in_first_sentence
is_title	similar word suffixes	
is_digit		
POS tag		
POS tag suffix		
is_first_word		
is_last_word		

Table 2: Features used in CRF experiments.

contains the sequence of words *mistakenly amputated my left leg*. Second, unlike relations or entities, it is non-obvious how to classify negligence related evidence into categories. This becomes problematic when generalising the extraction rules from one domain to another. To the best of our knowledge, the problem of extracting medical negligence related evidence from free text data has not been studied before.

3 Evidence Extraction

CRFs and LSTMs are two classes of models that perform well, and are often employed, in a range of sequence labelling tasks [Huang *et al.*, 2015; SHI *et al.*, 2015; McCallum and Li, 2003]. Both model classes are able to leverage historical and future sequence information when classifying the current sequence element. This makes them well suited to natural language processing tasks. One advantage LSTMs have over CRFs is their ability to learn feature representations that are specific to the task at hand. We employ both model classes in this work and compare their performance in the task of identifying negligent acts and consequential negative patient outcomes from medical negligence claim texts.

The dataset used in this evaluation comprises 2014 medical negligence claim summary texts. These texts contain statements describing negligent acts as well as any consequential negative patient outcomes (Figure 1). The texts were annotated with BIO tags delineating negligent act statements and consequential negative patient outcome statements by a domain expert working in medical negligence law. Table 1 shows some dataset statistics. Due to the confidential nature of the dataset, we are unable to share it.

4 Experiments

CRF models were trained using various combinations of the features listed in Table 2. The features listed in the left-hand column are common to most text tagging tasks. Those listed in the middle column were introduced to address the

LSTM settings
LSTM
LSTM + Glove
LSTM + Char
BiLSTM
BiLSTM + Char
BiLSTM + Glove + Char

Table 3: LSTM configurations used in these experiments.

problem of data sparseness. The similar word features require further explanation; these were generated using pre-trained Glove [Pennington *et al.*, 2014] embeddings: given a word, the N words with the highest cosine similarity were included as additional features; the value for N was varied ($N=\{1..10\}$). Similar word suffix features were also experimented with. The features in the right-hand column are domain specific. For example, it was observed that negligent act statements are often present in the first sentence of a claim text. Also, negligent act statements frequently contain medical terminology. The listed features were computed for each token in each sequence as well as the preceding and following tokens. All CRF models were trained using the sklearn-crfsuite Python package [Korobov, 2017]. The following hyper-parameters were tuned using a randomised search over 50 iterations: the Elastic net regularisation coefficient, the minimum feature frequency, and the possible state and transition features.

We experimented with various LSTM configurations (see Table 3). The baseline LSTM comprised a 50-dimensional word embedding input, a single LSTM layer of 16 hidden units, and a softmax output. This model was trained both with random and pre-trained Glove word embedding initialisation. Bi-directional LSTMs were also experimented with. The baseline model was extended to include character-level features. This was achieved using a convolutional layer containing 8 hidden units, with a 16-dimensional character embedding input. All LSTM models were trained using the NCRF++ Python package [Yang and Zhang, 2018]. Each LSTM was trained for 100 epochs using stochastic gradient descent with a learning rate of 0.015, a learning rate decay of 0.05, and a batch size of 32. During training, each model was evaluated using a validation set at the end of each epoch, and the best performing model (across the 100 epochs) was selected for use in the evaluation. Training was repeated 5 times for each LSTM configuration in order to reduce the influence of pathological local minima, but none were observed, therefore we randomly selected one of the 5 models for the evaluation (for each of the different configurations).

5 Results

The different methods were compared using a 5-Fold Cross Validation scheme. Performance metrics were computed both at the sequence level and the token level. Token level metrics were computed using the negligent act and negative patient outcome labels only (i.e. “other” tags were ignored). Neither evaluation scheme is perfectly suited to identifying the best

CRF feature set	Prec	Rec	F1
Base	0.486	0.385	0.428
Base + stem	0.486	0.384	0.427
Base + stem + suffix	0.492	0.382	0.429
Base + sentiment	0.487	0.378	0.424
Base + in_medical_lexicon	0.468	0.379	0.417*
Base + in_first_sentence	0.495	0.396	0.438
Base + 7 similar words	0.497	0.406	0.445*
Base + 6 similar words + suffix	0.489	0.406	0.443*

Table 4: Selected CRF model performance evaluated at the sequence level (micro-averaged). Note: Base refers to the the baseline CRF, which made use of generic features only; best results in bold font; * indicates a significant result ($P=0.05$, Bonferroni corrected).

Configuration	Prec	Rec	F1
LSTM	0.245	0.252	0.248
LSTM + Glove	0.195*	0.215*	0.205*
LSTM + Char	0.260	0.286	0.272
BiLSTM	0.230	0.242	0.236*
BiLSTM + Char	0.256	0.273*	0.264
BiLSTM + Char + Glove	0.197*	0.219*	0.207*

Table 5: LSTM performance evaluated at the sequence level (micro-averaged). Note: best results in bold font; * indicates a significant result ($P=0.05$, Bonferroni corrected).

performing sequence tagger. For example, evaluating models at the sequence level only will discount any examples where the system correctly identifies the vast majority of a sequence, but misses a single, minimally important term. Similarly, token level evaluation is imperfect as it can mask pathological behaviour. For example, a system can correctly identify the majority of a phrase but fail to identify a single important component (e.g. “no longer have any mobility in my”) and still score highly using this scheme. While it is not perfect, we suggest the phrase level evaluation is likely to be a better indicator of a model’s usefulness in practice. In order to test for the statistical significance of the results, we employed the corrected re-sampled t-test [Nadeau and Bengio, 2001], coupled with the Bonferroni correction for multiple comparisons [Dunn, 1961].

Table 6 compares the best performing CRF and LSTM models. The CRF model performed significantly better at the sequence level, while the LSTM offered significantly better token level performance. Inspecting extractions performed on a test set can be useful in comparing models. Figure 2 shows some example extractions performed using these two models. The outputs of the different models vary considerably: the two approaches only fully agree on a single instance (12 instances in total). The LSTM repeatedly fails to identify the beginning of the sequences: it only outputs a single **B** tag (a **B** tag indicates the first term in a sequence) out of a possible 12, whereas the CRF outputs 9 **B** tags. The LSTM exhibits additional undesirable behaviour: it erroneously splits sequences in two, often dropping a common word. It appears that the LSTM is giving too much consideration to the current word,

University hospital mistakenly amputated my left leg despite the fact the cancer was confined within my right leg. I will now need to undergo another leg amputation and will be confined to a wheelchair for the rest of my life.

I believe the University pharmacy to be negligent as they misprescribed me with ibuprofen when they should have given me paracetamol. I felt sick for a week as a result.

I believe the midwife at University hospital was at fault because she dropped my newborn Son. This caused his arm to break, and his head is now misshapen. We are unsure if his head will ever regain its original shape, or if he will have lasting problems with his arm.

I believe the GP at the Village Health Centre should have noticed the lump when I first presented with my symptoms. My cancer diagnosis has now been delayed by 15 months, and the prognosis is much worse.

Figure 2: Example extractions performed by the CRF and LSTM. Tokens underlined with blue were identified by the CRF only; tokens underlined with red were identified by the LSTM only, and tokens underlined with violet were identified by both models.

and the previous sequence information is discounted. Both approaches make some subtle mistakes that produce extractions that appear to be correct at a first glance, but are actually incorrect. For example, in the third example in Figure 2, the CRF identifies the sequence “lasting problems with his arm”, when in reality in the statement the author suggests they are unsure whether the child will have lasting problems with their arm. Extractions like this could prove to be problematic, if such a system is used to quickly extract the key case facts from a statement.

Tables 4 and 5 compare the different CRF feature sets and LSTM configurations. The different LSTM configurations performed similarly well, except for in cases where the word embeddings were initialised using pre-trained Glove vectors – in these instances the models performed significantly worse than the baseline LSTM. We also found that training a BiLSTM with character level features significantly improved recall. Moreover, we found that adding sparseness-counteracting features improved CRF performance. The best performing CRF model made use of similar word features (N=7). We also found adding domain specific features to be helpful: including whether or not a word occurs in the claim text’s first sentence as a feature significantly improved token level performance. This feature was strongly associated with the negligent act class.

6 Discussion and Conclusion

In this set of experiments we found both CRF and LSTM models were able to extract litigation-relevant information from medical negligence claim texts. We observed that the CRF was better able to identify entire useful phrases, while

Sequence level evaluation						
	Prec		Rec		F1	
	CRF	LSTM	CRF	LSTM	CRF	LSTM
NA	0.50*	0.32	0.43	0.41	0.46*	0.36
O	0.49*	0.22	0.39*	0.23	0.44*	0.23
AVG	0.49*	0.26	0.40*	0.29	0.44*	0.27

Token level evaluation						
	Prec		Rec		F1	
	CRF	LSTM	CRF	LSTM	CRF	LSTM
B-NA	0.68	0.87*	0.59	0.60	0.63	0.71*
I-NA	0.63	0.81*	0.67	0.77*	0.65	0.79*
B-O	0.60	0.75*	0.48*	0.24	0.54*	0.37
I-O	0.62	0.72*	0.55*	0.50	0.59	0.59
AVG	0.63	0.78*	0.60	0.61	0.61	0.67*

Table 6: Comparison of the best performing CRF and LSTM models evaluated at the phrase and token levels. Note: “NA” refers to negligent act; “O” refers to consequential negative outcome; AVG refers to the micro average; best results in bold font; * indicates a significant result (P=0.05, Bonferroni corrected).

the LSTM was able to assign labels to tokens with higher precision. The best performing CRF model’s ability to identify evidence is likely sufficient for it to be useful in practice. We found that enriching the CRF features with similar words, computed using pre-trained word embeddings, improved the CRF’s performance. We also observed including domain specific features improved the CRF’s performance. While the evaluation suggests the CRF is better suited to this task than the LSTM, we recognise it may well be biased in favour of the CRF. This is because we experimented with few LSTM architectures, and the architecture is an important hyperparameter when training neural network models. In future work we plan to experiment further with the LSTM architecture. Specifically, we plan to vary the dimensionality of the various embedding and hidden layers. We also plan to experiment with a CRF output layer with the view that this may improve the LSTM’s sequence level performance. We also plan to collect more data, which may benefit both approaches and further assist with the development of our automated tools for processing medical negligence documents.

References

- [Bollegala *et al.*, 2018] Danushka Bollegala, Simon Maskell, Richard Sloane, Joanna Hajne, and Munir Pirmohamed. Causality patterns for detecting adverse drug reactions from social media: Text mining approach. *JMIR Public Health and Surveillance*, 4(2):e51, May 2018.
- [Dunn, 1961] Olive Jean Dunn. Multiple comparisons among means. *American Statistical Association*, pages 52–64, 1961.

- [Guo *et al.*, 2009] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *SIGIR 2009*, pages 267–274, 2009.
- [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv e-prints*, page arXiv:1508.01991, Aug 2015.
- [Korobov, 2017] Mikhail Korobov. sklearn-crfsuite. <https://github.com/TeamHG-Memex/sklearn-crfsuite>, 2017.
- [Kuru *et al.*, 2016] Onur Kuru, Ozan Arkan Can, and Deniz Yuret. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.
- [Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML 2001*, pages 282–289, 2001.
- [Mandya *et al.*, 2017] Angrosh Mandya, Danushka Bollegala, Frans Coenen, and Katie Atkinson. Frame-based semantic patterns for relation extraction. In *Proc. of the 15th International Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 51–62, 2017.
- [McCallum and Li, 2003] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
- [Miwa and Bansal, 2016] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [Nadeau and Bengio, 2001] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 2001.
- [NHS, 2018] NHS. Annual report and accounts 2017/18. Technical report, National Health Service (NHS) Resolution, 2018.
- [Pennington *et al.*, 2014] Jeffery Pennington, Richard Socher, and Christopher D. Manning. Glove: global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.
- [Ritter *et al.*, 2011] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP'11*, pages 1524 – 1534, 2011.
- [Rud *et al.*, 2011] Stefan Rud, Massimiliano Ciaramita, Jens Muller, and Hinrich Schütze. Piggyback: Using search engines for robust cross-domain named entity recognition. In *ACL'11*, 2011.
- [Shen *et al.*, 2018] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, 2018.
- [SHI *et al.*, 2015] Xingjian SHI, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 802–810. Curran Associates, Inc., 2015.
- [Sloane *et al.*, 2015] Richard Sloane, Orod Osanlou, David Lewis, Danushka Bollegala, Simon Maskell, and Munir Pirmohamed. Social media and pharmacovigilance: A review of the opportunities and challenges. *British Journal of Clinical Pharmacology*, pages 910–920, 2015.
- [Toyabe, 2012] Shin-ichi Toyabe. Detecting inpatient falls by using natural language processing of electronic medical records. *BMC Health Services Research*, 12(1), Dec 2012.
- [Yang and Zhang, 2018] Jie Yang and Yue Zhang. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.