# MedCATTrainer: A Biomedical Free Text Annotation Interface with Active Learning and Research Use Case Specific Customisation

**Thomas Searle**[1], **Zeljko Kraljevic**[1], **Rebecca Bendayan**[1], **Daniel Bean**[1], **Richard Dobson**[1,2]
[1]Department of Biostatistics and Health Informatics, Kings College London, London, U.K.
[2]Institute of Health Informatics, University College London, 222 Euston Road, London NW1 2DA, U.K.
{firstname.lastname}@kcl.ac.uk

## Abstract

We present MedCATTrainer[1] an interface for building, improving and customising a given Named Entity Recognition and Linking (NER+L) model for biomedical domain text. NER+L is often used as a first step in deriving value from clinical text. Collecting labelled data for training models is difficult due to the need for specialist domain knowledge. MedCATTrainer offers an interactive web-interface to inspect and improve recognised entities from an underlying NER+L model via active learning. Secondary use of data for clinical research often has task and context specific criteria. MedCATTrainer provides a further interface to define and collect supervised learning training data for researcher specific use cases. Initial results suggest our approach allows for efficient and accurate collection of research use case specific training data.

## 1 Introduction

We present a flexible web-based open-source use-case configurable interface and workflow for biomedical text concept annotation - MedCATTrainer[2].

Murdoch and Detsky (2013) estimates that 80% of biomedical data is stored in unstructured text such as Electronic health records (EHRs). Although EHRs have seen widespread global adoption, effective secondary use of the data remains difficult (Elkin et al., 2010). However, significant progress has been made on agreement and usage of standardised terminologies such as the Systematized Nomenclature of Medical Clinical Terms (SNOMED-CT) (Stearns et al., 2001) and the Unified Medical Language System (UMLS)(Bodenreider, 2004). Annotating EHR text with these concept databases is often seen as a first step in delivering data driven applications such as precision medicine, clinical decision support or real time disease surveillance (Assale et al., 2019).

EHR text annotation is challenging due to the use of domain specific terms, abbreviations, misspellings and terseness. Text can also be 'copy-pasted' from prior notes, structured tables entered into unstructured form, content with varying temporality and scanned images of physical documents (Botsis et al., 2010). Annotation is further complicated as researchers have task and context specific parameters. For example, whether family history or suspected diagnoses are considered relevant to the task.

MedCAT[3], manuscript in preparation (Zeljko and Lucasz, 2019), is a **Med**ical **C**oncept **A**nnotation **T**ool that uses unsupervised machine learning to recognise and link medical concepts with clinical terminologies such as UMLS. MedCAT, like similar tools, uses a concept database to find and link concept mentions inside of biomedical documents. In addition it has disambiguation, spell-checking and the option for supervised learning for improved disambiguation.

We introduce a novel web based application that supplements usage of a biomedical NER+L models, such as MedCAT. Our contributions are as follows:

1. **Concept Inspection and Addition:** an interface that to inspect the identified concepts from free text, and add missing concepts to an existing NER+L model. This interface aligns with MedCAT, but could also be used with other models that have similar capabilities.

2. **Active Learning:** an interface for active learning, enabling users to provide minimal

[1]https://www.youtube.com/watch?v=lM914DQjvSo
[2]https://github.com/CogStack/MedCATtrainer
[3]https://github.com/CogStack/MedCAT



*Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations)*, pages 139–144
Hong Kong, China, November 3 – 7, 2019. 

training data to assist in improving and correcting the NER+L. This interface requires that the backing NER+L system supports active learning.

3. **Clinical Research Question Specific Annotation:** a further interface for configurable use case specific annotation of identified concepts. Allowing for the collection of research question specific training data. For example, annotating specific temporal features of a concept.

## 2 Related Work

Outside of the biomedical domain general purpose annotation interfaces have been developed for most popular NLP tasks such as NER, NEL, relation extraction, entity normalisation, dependency parsing, chunking etc. Popular choices include open-source tools such as BRAT (Stenetorp et al., 2012) that also allows for managing the distribution, monitoring and collection of annotated corpora. General purpose tools with active learning include the commercial product Prodigy[4]. Although these tools are mature and offer advanced features they can be complex to setup and do not offer integration with existing biomedical domain NER+L systems.

Prior work on biomedical NER+L includes MetaMAP (Aronson, 2001) and CTakes (Savova et al., 2010). Both have provided interfaces to inspect recognised entities but they have not provided means to correct and amend concepts or specify further annotations for specific research questions.

Another tool for biomedical NER+L, SemEHR Wu et al. (2018), offers features to add custom pre and post processing steps and research specific use cases, but does not directly improve the NER+L model via an interface. Instead it treats the provided NER+L model as a black-box model.

## 3 MedCATTrainer

MedCATTrainer is a web-based interface for inspecting, adding and correcting biomedical NER+L models through active learning. An additional interface allows for research specific annotations to be defined and collected for training of supervised learning models.

[4]https://explosion.ai/blog/prodigy-annotation-tool-active-learning

The interfaces are built with Vue.js[5] for the front-end and the python[6] web framework Django[7] for the web API and integration with NER+L models such as MedCAT. We use the Django admin features to allow administrators to configure research question specific supervised learning tasks.

MedCATTrainer is deployed via a Docker[8] container. This ensures users can build, deploy and run MedCATTrainer cross-platform without lengthy build and run processes, advanced infrastructure knowledge or root access to systems. This is especially important in health informatics as hospital infrastructure is often restrictive. MedCATTrainer allows researchers to build on top of existing biomedical domain ontologies, such as UMLS, for two use cases. Firstly, improving the underlying NER+L model by adding synonyms, abbreviations, multi-token concepts and misspellings directly from the interface. Secondly, by allowing research use case specific annotations to be defined and collected for training of supervised learning models.

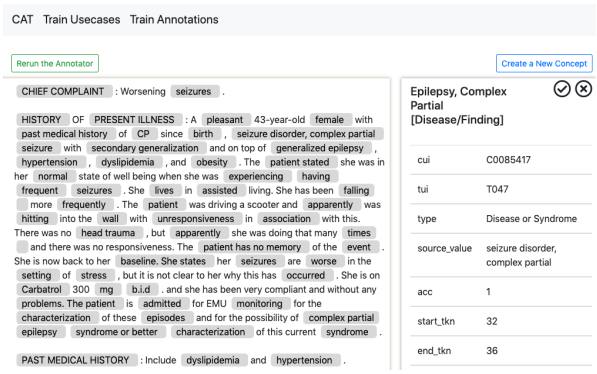
### 3.1 Concept Inspection and Addition

Figure 1a shows the 'Train Annotations' interface. Users can inspect and correct the concepts identified by the underlying NER+L model. Entities that have not been recognised can also be added to the NER+L model concept database. This allows researchers to test the learnt entity recognition/linking capabilities of the model whilst tailoring it to recognise sub-domain specific lexicon. This can include abbreviations or misspellings common to specific corpora. Figure 1b shows the form entry to add new concepts to the underlying concept database. Semantically equivalent texts can be added under the same Concept Unique Identifier along with synonyms. Advanced NER+L tools (e.g. MedCAT) learn from the contextual embeddings of words to disambiguate future occurrences. MedCATTrainer provides a textbox for entering the surrounding context tokens to assist with concept disambiguation.

### 3.2 Active Learning

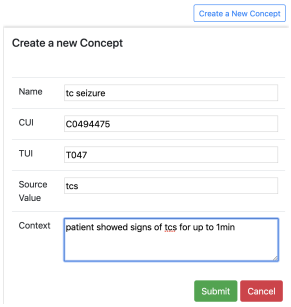
Annotating biomedical domain text for NER+L requires expert knowledge and therefore cannot be

[5]https://vuejs.org/
[6]https://www.python.org/
[7]https://www.djangoproject.com/
[8]https://www.docker.com




(a) The MedCATTrainer interface for viewing identified concepts by the underlying NER+L model of a publicly available[a] neurological consultation summary showing the concept metadata and active learning feedback input controls.

[a] https://bit.ly/2RLcdJx




(b) Side panel for the addition of new concepts.

Figure 1: The interfaces for inspecting annotations and the addition of concepts.

easily crowd sourced. Active learning is a common approach to provide a minimal set of high value training examples for manual annotation. Examples are valued with respect to expected improvement in classification performance once labelled and the model retrained (Settles, 2009).

We use a simple strategy of certainty based selective sampling (Lewis and Catlett, 1994) to display low confidence examples. Concretely, given a trained model $\mathbf{M}$, and the total set of annotations predicted on a new document $d$ by model $\mathbf{M}$ is $\mathbf{L} = \{l_1, l_2, \ldots l_n\}$ where the model labelled the document with $n$ annotations. An annotation $l_i$ has an associated confidence $c_{l_i}$ probability in the annotation. An annotation manager defines $\delta$, a confidence cutoff score. The set of annotations $\mathbf{A}$ shown to an annotator is therefore $\Phi(\mathbf{L})$ where $\Phi(l_i) = c_{l_i} > \delta$.

Each human annotator is instructed to review each identified concept and provide feedback on correctness. Feedback is provided through the action of clicking the 'tick' for correct or 'cross' for incorrect as shown in the top right of Figure 1a.

If an identified concept is incorrect human annotators are asked to provide feedback, rerun the NER+L model (top left 'Rerun the Annotator'), and then confirm if the misidentified concept has been corrected. More feedback can be provided if needed. Our pilot test users found this quickly resulted in the correctly identified and linked concept as text spans often only have one or two alternative concepts.

### 3.3 Clinical Research Question Specific Annotation

It would be infeasible to have a clinical terminology to define every possible contextual representation of a concept. For example, disambiguation of 'seizure' for a symptom of epilepsy and 'first seizure clinic' for a clinic that provides epilepsy care or 'history of seizures' for a historical case of epilepsy.

Our second interface solves this problem by allowing clinical researchers to define use case orientated tasks and associated annotations for previously identified and linked concepts. Custom classifiers are then trained and layered over the existing NER+L model for context specific concept disambiguation. An example configured screen for 'Temporality' and 'Phenotyping' tasks for an ongoing clinical research project is shown in Figure 2 - using replacement publicly available data. The top bar lists the overall task name followed by the number of documents to be annotated. The top right corner opens the current task help document, listing annotation guidelines for this use-case.

The left panel itemises each text span, the associated Concept Unique Identifier (CUI) - that the NER+L model has identified and linked with the text, and the current value of each task specific annotation. The value 'n/a' indicates the task has not been completed for that span. Users can choose any order of the text spans to annotate. The currently selected text span is highlighted in the table and within the central text area showing the entirety of the document. Clinical notes can be

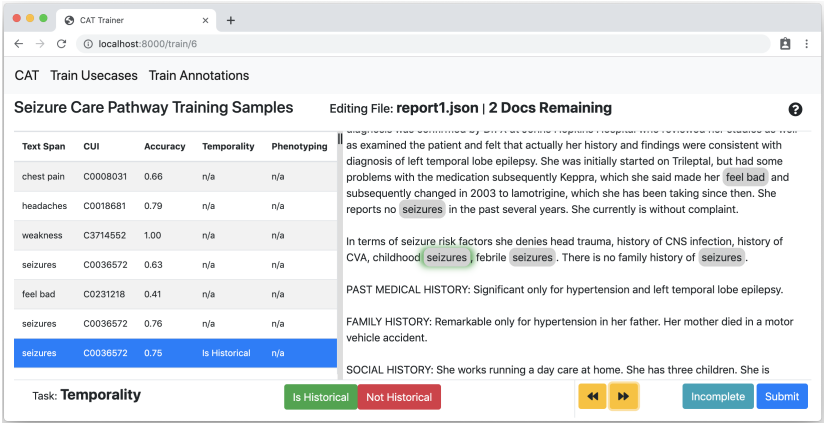



Figure 2: Task and context specific annotation interface configured for 'Temporality' and 'Phenotype' tasks

long in length. Clicking a text span from the sidebar scrolls the central text area to the corresponding span assisting human annotators in locating the span to annotate. The text area also highlights each spans current annotated value for the current task.

The bottom bottom bar lists the current task and the possible annotation values. Figure 2 shows the 'Temporality' task and the associated annotation values 'Is Historical' and 'Not Historical'. The values are in context to a seizure care pathway use case and are defined as any currently experienced mention of seizure symptoms in present clinical encounter. Use cases and associated tasks values are configurable via the admin interface.

The bottom right corner provides navigation between text spans and tasks via the arrow buttons. Navigating between spans highlights the current span to be annotated in the main left sidebar and auto scrolls to the next span in the main text area. The navigation controls here, the sidebar and the main text area allow human annotators to complete the task in any order they are comfortable.

The 'Incomplete' button marks the current document to be revisited at a later date. Samples are marked incomplete if the NER+L model has misidentified the concept or there is a genuine ambiguity. The 'Submit' button marks the document as complete. Both actions store and retrieve the next document if there is one available. If there are no more files to annotate a dialog prompts the user to return to the home screen.

Corpora are currently directly uploaded via a use case management screen. Future deployments will directly ingest documents via an elasticsearch[9] connector to hospital EHR deployments of CogStack (Jackson et al., 2018) an EHR ingestion, transformation and search service deployed at King's College Hospital (KCH) and South London and Maudsley(SLaM) NHS Foundation Trusts, UK.

## 4 Results

We ran an initial small scale pilot experiment to test the suitability of our use case specific tool to quickly and accurately collect training data labelling the temporal features of seizure symptoms. This is similar to the task shown in Figure 2. We used MIMIC3 (Johnson et al., 2016), a de-identified publicly available database of ICU admission data that includes observations, consultation and discharge summary reports. We randomly sampled 127 discharge summaries that contained one or more token occurrences that match the regular expression 'seizure|seizre|seizur|siezure', where | is an OR operator between the text tested to be present. We intentionally rely on a rule-based NER mode (i.e. the regex) here to demonstrate our tools flexibility to use possible alternatives to MedCAT if desired.

We asked 2 human non-clinical annotators to label temporal features of each occurrence in relation to a 'present', i.e. 'chief complaint: seizure' or 'historical', i.e. 'family history of seizures', mention of the term. Both took approximately 35 minutes to review all 127 documents. We achieve an percent agreement of 89% and a Cohen's Kappa $\kappa = 0.695$, Table 1. Both annotators marked some records as incomplete as they either mostly referred to non symptomatic mentions

[9] https://www.elastic.co/

| | R1* | R2* | R1 | R2 |
|---|---|---|---|---|
| # Documents | 107 | 117 | 100 | 100 |
| # Concepts | 351 | 344 | 317 | 317 |
| # Historical | 67 | 80 | 79 | 65 |
| # Not Historical | 276 | 264 | 238 | 252 |

Table 1: Total labelled 'seizure' symptom concepts and for each human annotator (R1, R2) for the 'temporality' task of labelling concepts that have occurred the past relative to the hospital episode. * indicates raw numbers before taking into account the intersection of notes between annotators

of seizure, i.e. 'anti-seizure meds prophylaxis' or the prevention of future seizures. This resulted in each rater having differing total documents 'submitted' as there are some document with mixes of the above occurrences. We took the intersection of submitted documents from both raters to compute the final agreement scores.

Using the collected data we fit a simple Sckit-learn[10] Random Forest (RF) classifier model demonstrating the effectiveness of the data collection in being able to easily fit a well performing model for the task of recognising temporality of seizure symptoms. We took a random 70/30 train test split, took 100 characters either side of the labelled 'seizure' occurrence, tokenized the plain text on whitespace then used a TF-IDF vectoriser with the default English stop-words list. We ran a grid search across TF-IDF and random forest classifier parameters, with a 3 fold cross validation and found the best fitting parameters: TF-IDF features 500 (range:500, 1000, 10000), RF maximum number trees of 100 range(100, 300, 500, 1000) and maximum tree depth 20 (range: 5, 20, 50, 75). We achieve an accuracy of this binary classification task of 92% and f1 score .79.

## 5 Discussion and Future Work

From our labelling exercise we demonstrate the speed and accuracy of our configurable use case specific interface. Strong scores across % agreement, Cohen's Kappa and trained model accuracy indicate good agreement between annotators, interpretations of the task and reasonable signal captured even with this small data set. Although, it is likely the model is over-fitting due to the size of the data set. Given the prior experiment - across two raters - gathering enough accurate data to, for example, fine-tune a pretrained language model based classifier would be of the order of hours of manual labelling for approx 2k samples. We see this rapid labelling ability as a key strength of our interface.

We foresee that trained classifiers will likely generalise to additional research questions. For example language used to express temporality of seizures is likely to be similar to temporality of stroke or myocardial infarction.

Generally, training models across use cases will likely capture shared semantics. This suggests particular use cases would require less examples to train as annotated data or the model itself could be reused, therefore jump-starting clinical research. If a model is not performing for a new use case, further data could be collected to fine tune the model to a specific task, context or sub-domain corpora.

Clinically, domain experts in the neurology department of KCH, with varying levels of expertise (medical student to practising consultant) are scheduled to participate in the use case shown in Figure 2 in the coming months.

Our initial testing, not shown above due to space, of the active learning approach for improving the bound NER+L model suggests we can improve performance with minimal training data.

## 6 Conclusions

We have presented a lightweight, flexible, web-based, open-source annotation interface for biomedical domain text. MedCATTrainer is integrated with a biomedical NER+L model and allows for addition of missing concepts, improvements to the underlying NER+L model through active learning, and a configurable interface for clinical researchers to define annotations specific for their research questions. Preliminary results show promise for our interface and our approach to biomedical NER+L, which is often seen as a first step in deriving value from data sources such as electronic health records.

## Acknowledgments

DMB is funded by a UKRI Innovation Fellowship as part of Health Data Research UK (MR/S00310X/1). RB is funded in part by grant MR/R016372/1 for the Kings College London MRC Skills Development Fellowship programme funded by the UK Medical Research Council

[10] https://scikit-learn.org/stable/index.html

(MRC) and by grant IS-BRC-1215-20018 for the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. RD's work is supported by 1.National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. 2. Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust. 3. The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. The views expressed are those of the author(s) and not necessarily those of the NHS, MRC, NIHR or the Department of Health and Social Care.

## References

A R Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, pages 17–21.

Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. 2019. The revival of the notes field: Leveraging the unstructured content in electronic health records. *Front. Med.*, 6:66.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70.

Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. Secondary use of EHR: Data quality issues and informatics opportunities. *Summit Transl Bioinform*, 2010:1–5.

Peter L Elkin, Brett E Trusko, Ross Koppel, Ted Speroff, Daniel Mohrer, Saoussen Sakji, Inna Gurewitz, Mark Tuttle, and Steven H Brown. 2010. Secondary use of clinical data. *Stud. Health Technol. Inform.*, 155:14–29.

Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. 2018. CogStack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC Med. Inform. Decis. Mak.*, 18(1):47.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In William W Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA).

Travis B Murdoch and Allan S Detsky. 2013. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352.

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. *J. Am. Med. Inform. Assoc.*, 17(5):507–513.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

M Q Stearns, C Price, K A Spackman, and A Y Wang. 2001. SNOMED clinical terms: overview of the development process and project status. *Proc. AMIA Symp.*, pages 662–666.

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*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson, Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart, and Richard J B Dobson. 2018. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.*, 25(5):530–537.

Kraljevic Zeljko and Roguski Lucasz. 2019. Cogstack/medcat: First release of medcat.