Detection of Medical Events in Discharge Summaries

Adyasha Maharana, Manasa Bollavaram

March 19, 2017

1 Introduction

Physicians spend considerable time in assimilating information about a patient's condition. Aggregation of medical information from multiple sources and its arrangement into a neat clinical timeline, can save precious time. At the end of a hospital stay, every patient is issued a 'Discharge Summary' which contains critical information about the patient's condition. The subsections of a discharge summary include 'Admission Date', 'Discharge Date', 'History of Present Illness' and 'Hospital Course' among others. Information is primarily expressed in free text containing medical shorthands, clinical terms and informal grammatical constructs. Temporally sorted data extracted from these summaries can also have downstream use in inference and prediction of clinical conditions.

The task of extraction of clinical timeline belongs to the sub-field of Information Extraction in Natural Language Processing. The first part of this system is a Named Entity Recognition sub-system which is trained to identify various types of medical events from a clinical note, and classify its modality and polarity. The major bottlenecks of this system are limited, expensive annotated data and lack of generality. Traditionally, this task has been approached as a structured inference problem, which is very well-suited to small datasets. On the downside, it takes a good amount of domain expertise and lots of time to develop a high-performing Medical Event Detection system based on structured inference. The model also runs the risk of overfitting to a specific medical domain depending on the distribution of training data. The i2b2 center has been promoting work in this field through Shared-Task challenges and openly available datasets. The effort to build an end-to-end Temporal Evaluation system has been continued by the annual SemEval workshop through the Clinical TempEval task. The latest challenge in this series of tasks was to develop a system trained on discharge summaries of colon cancer patients, to make predictions for brain cancer patients. [1]

For this project, we will focus on developing a robust and generic Medical Event Detection system. We intend to circumvent hand-crafted feature engineering by using deep neural network architectures and build on top of unsupervised methods to make use of unannotated clinical data.

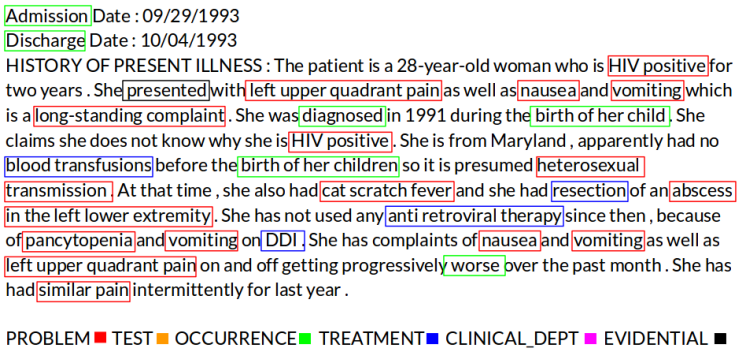

Figure 1: Example of a Discharge Summary with Annotation for Medical Events

# Phrases	Training	Development	Test
Sentences	3825	547	1092
PROBLEM	3207	469	973
TEST	1684	253	456
TREATMENT	2545	343	659
OCCURRENCE	2061	341	649
CLINICAL DEPT	611	96	180
EVIDENTIAL	475	78	126
Total Events	10583	1580	3043

Table 1: Summary of Dataset

2 Dataset and Annotation

This problem statement was posed by the 2012 i2b2 Challenge as a Shared NLP Task. [6] A set of 310 de-identified and annotated discharge summaries taken from two medical centers, was released as dataset for this task and is currently available on their website for research purposes. The same dataset will be used for this project. A typical discharge summary from the dataset with event annotations looks as shown in Figure 1.

Extraction of clinical timeline is accomplished in three steps: Event Detection, Temporal Phrase Detection and Temporal Relation Identification. Accordingly, the annotation has been performed for three categories: Event, Temporal Expression and Temporal Relation. For this project, we work only on the Event annotations. Summary of the dataset and event annotations is presented in Table 1.

3 Literature Review

The most successful attempts at building a medical event detection system have combined machine learning and rules to come up with a hybrid system that consists of a supervised training module and heuristic rules for post-processing. [3] The field of Medicine adds a dimension of difficulty to this NER task, because specialized knowledge is required to recognize clinical/biomedical terms such as 'myocardial infarction' or 'benign tumor'. The most used public knowledge base is the UMLS Metathesaurus, an integral component of biomedical informatics systems. [2] Feature-rich hybrid systems combined with knowledge bases have been able to achieve high performance with F1 score of up to 0.9. [7] Private knowledge bases have been substituted with (pain-staking) pattern-matching, word clustering features, and rules in a system that achieves very high precision and F1 score of 0.92. [10] These systems have been tested on the 2012 i2b2 dataset. We don't know the extent to which this dataset is representative of the entire clinical domain; hence, it is difficult to comment on the generality of such systems.

The website for recently concluded Clinical TempEval Task 2017 enlists scores of the best-performing systems; the leaderboard of event detection stands at F1-score of 0.72 and 0.76 for unsupervised and supervised domain adaptation respectively. Details about workings of the systems are yet to be released. However, a mere 4 point gap between unsupervised and supervised systems shows that, in spite of lack of sufficient data, unsupervised methods can perform reasonably well in developing generic medical event detection systems.

Before the advent of deep learning, Conditional Random Field (CRF) had been the most popular statistical modelling algorithm for structured prediction. With recent innovations, there has been a surge in the development of neural network architectures for superior sequence modelling and automatic feature engineering. Bi-directional LSTM-RNNs have been successfully deployed for NER in various languages. [2] More recently, word embeddings and RNNs have found use in the biomedical domain for tasks ranging from sequence tagging to diagnosis from time-series clinical measurements. [5] [8]

4 Method

4.1 Structured Inference

The current state-of-art systems for end-to-end temporal evaluation consist of medical event detection sub-systems based on Conditional Random Fields. [7] [9] We attempt to replicate these systems as far as possible, to establish a baseline. Various modules are used in this system such as, POS Tagger trained on medical text, statistical section chunker etc. Orthographic, syntactic and semantic features are designed and tested for various lengths of context window. An overview of the final system is presented in Figure 2.

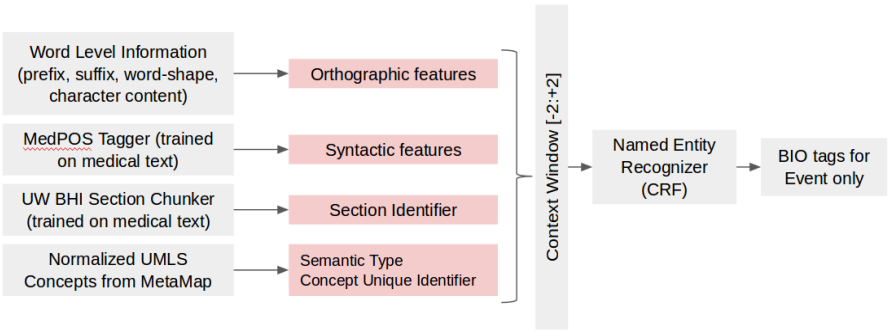

Figure 2: Structured Inference Model

4.2 Recurrent Neural Networks

Currently, the state-of-art in deep NER architecture is a combination of Bi-directional LSTM-RNN and a layer of Conditional Random Field. [4] The vectors used for training RNN are concatenations of character-level and word-level embeddings. Character-LSTMs are used to represent orthographic features, which are significant sources of information for NER. Bi-directional RNNs conveniently capture information from start and end of the word (prefi, suffix) and to left and right of the word (context window). Bi-directional LSTMs have been efficient in other sequence modelling tasks such as POS tagging, but their success in NER has been limited. CRFs promote joint reasoning on event tags and amplify the contribution of transition probabilities to the final tagging decision.

We implement this architecture for medical-event detection. For the lack of publicly available clinical pre-trained embeddings, we run our model with randomly initiated embeddings in the first phase. Further, we tune the architecture for our task by experimenting with different word, hidden layer dimensions.

5 Results & Discussion

5.1 Conditional Random Field

The CRF model achieves a 5-fold cross-validation precision, recall and F1 score of 0.75, 0.72 and 0.73 respectively, More fine-grained results are reported along with results from RNN for easier comparison. Some of the most significant features included prefixes and suffixes like 'cin', 'tics' for 'TREATMENT', 'repo' for evidential; UMLS semantic type being 'Disease/Finding' of current and previous words for 'PROBLEM', 'TEST' etc.

Precision	Recall	F1 Score	Support
0.75	0.72	0.73	7355

Table 2: Aggregate Performance of Conditional Random Field

Model	Precision	Recall	F1 Score
Default	0.66	0.68	0.67
Character Embedding Dimension = 50	0.71	0.63	0.67
Character LSTM Dimension = 50	0.73	0.67	**0.70**
Word Embedding Dimension = 200	0.73	0.65	**0.69**
Word LSTM Dimension = 200	0.68	0.67	0.68
Dropout = 0.3	0.69	0.66	0.68

Table 3: Aggregate Performance of Variations of RNN+CRF on Validation set

5.2 Bi-directional LSTMs & CRF

For hyperparameter tuning, various configurations of the architecture were run in addition to the default setting. The parameters for default setting were as follows:

- Character Embedding Dimension: 25
- Character LSTM Hidden Layer Dimension: 25
- Word Embedding Dimension: 100
- Word LSTM hidden Layer Dimension: 100
- Optimization Method and Initial Learning Rate: Stochastic Gradient Descent, 0.05
- Dropout: 0.5

In the tuning phase, models were run for 20 epochs. Improvement was seen when the character hidden layer dimension was raised to 50 and word embedding dimension was doubled, in two individual runs. Other modifications did not seem to affect the default performance by a large margin. For the final run, the character hidden layer dimension was 50 and default configuration was retained for all other parameters. The dimensions were not increased beyond these values because it would lead to a simultaneous increase in number of parameters. Considering the limited amount of dataset we have, the model would be at a higher risk of overfitting in that case.

CRF fares better than the final Neural Network model for detecting PROBLEM, TEST and TREATMENT. However, the Neural Network is able to detect CLINICAL_DEPT, OCCURRENCE and EVIDENTIAL entities better. A possible reason for this behavior might be that semantic information from UMLS boosts the structured inference model's

Tag	Precision	Recall	F1
PROBLEM	0.67	0.69	0.68
TEST	0.78	0.70	0.74
TREATMENT	0.73	0.72	0.72
OCCURRENCE	0.69	0.53	0.60
CLINICAL DEPT	0.86	0.57	0.68
EVIDENTIAL	0.84	0.62	0.71

Table 4: Results for Bi-directional-LSTM-CRF Model after 25 epochs

Tag	Precision	Recall	F1
PROBLEM	0.75	0.78	0.77
TEST	0.82	0.71	0.76
TREATMENT	0.73	0.74	0.73
OCCURRENCE	0.60	0.48	0.53
CLINICAL DEPT	0.84	0.83	0.84
EVIDENTIAL	0.57	0.52	0.54

Table 5: Results for CRF Model

performance while the Neural Network is not able to leverage semantic concepts from the small corpus. The character embeddings seem to be having an upper hand over the hand-crafted orthographic features in structured inference.

6 Pre-trained Word Embeddings & Future Work

Pre-trained word embeddings have been shown to improve the performance of the Bi-directional-LSTM & CRF architecture by up to 7 points. With generic clinical text embeddings, it will also be possible to improve domain adaptation of the medical event detection system. A huge amount of unannotated clinical data is publicly available as the MIMIC-III Database. We ran word2vec on text from discharge summaries and examined the word clusterings. The MIMIC data embeddings represented much more fine-grained information and generic information than the embeddings from i2b2 dataset. Some of the clusters and their nearest neighbors trained from a subset of MIMIC data are given below:

- *admission*: arrival, discharge, transfer, admit, presentation, hospitalization, qd
- *dr*: drs, md, drlast, first, uta, transaminase, drname, doctor
- *right*: left, l, r, bilateral, rightsided, uta, leftsided, velcade
- *stable*: afebrile, unchanged, good, stabilized, asymptomatic, unremarkable, improved
- *status*: manifestations, exsanguinating, depletion, gog, impulsive, nephew, osteolysis

- *ct*: mri, cta, cxr, xray, imaging, ultrasound, noncontrast

Clearly, the model is learning to interpret medical abbreviations, clinical terms for occurrence, health condition etc. Closer examination of its t-SNE plot also shows clustering between surgical equipments, pain/fever symptoms, body organs, chronic diseases etc. On the other hand, i2b2 word clusterings are rough and vague, which could have been a major bottleneck for performance of this architecture for medical event detection. Some examples from i2b2 word embeddings are as follows:

- *admitted*: discharged, titrated, transferred, started, transported, referred, levo
- *right*: left, Restrictive, bilateral, uterus, thrombosis, losses, Afib, likelihood
- *chest*: vasculopathy, systemic, Chest, fevers, carry, instructed, Seen, renal

We intend to use MIMIC-III embeddings for future runs on this architecture.

References

[1] URL: `http://alt.qcri.org/semeval2017/task12/`.

[2] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology". In: *Nucleic acids research* 32.suppl 1 (2004), pp. D267–D270.

[3] Min Jiang et al. "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries". In: *Journal of the American Medical Informatics Association* 18.5 (2011), pp. 601–606.

[4] Guillaume Lample et al. "Neural architectures for named entity recognition". In: *arXiv preprint arXiv:1603.01360* (2016).

[5] Zachary C Lipton et al. "Learning to diagnose with LSTM recurrent neural networks". In: *arXiv preprint arXiv:1511.03677* (2015).

[6] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge". In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 806–813.

[7] Buzhou Tang et al. "A hybrid system for temporal information extraction from clinical text". In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 828–835.

[8] Yonghui Wu et al. "A study of neural word embeddings for named entity recognition in clinical text". In: *AMIA Annual Symposium Proceedings*. Vol. 2015. American Medical Informatics Association. 2015, p. 1326.

[9] Yan Xu et al. "An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge". In: *Journal of the American Medical Informatics Association* 20.5 (2013), pp. 849–858.

Figure 3: tSNE Projection of Word Embeddings learnt from MIMIC-III

[10] Yan Xu et al. “Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries”. In: *Journal of the American Medical Informatics Association* 19.5 (2012), pp. 824–832.