Relation Extraction from News Articles (RENA): A Tool for Epidemic Surveillance

Jaeff Hong[*2], **Duong Dung**[*1], **Danielle Hutchinson**[1], **Zubair Akhtar**[1], **Rosalie Chen**[1], **Rebecca Dawson**[1], **Aditya Joshi**[†2], **Samsung Lim**[1 3], **C Raina MacIntyre**[1] **and Deepti Gurdasani**[1 4 5],

[1]Kirby Institute, University of New South Wales, Sydney, Australia
[2]School of Computer Science & Engineering, University of New South Wales, Sydney, Australia
[3]School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia
[4]William Harvey Research Institute, Queen Mary University of London, London, UK
[5]School of Medicine, University of Western Australia, WA, Australia
jaeff.hong@gmail.com, {duong.dung, danielle.hutchinson, zubair.akhtar, rosalie.chen, rebecca.dawson, aditya.joshi, s.lim, r.macintyre, d.gurdasani}@unsw.edu.au

Abstract

Relation Extraction from News Articles (RENA) is a browser-based tool designed to extract key entities and their semantic relationships in English language news articles related to infectious diseases. Constructed using the React framework, this system presents users with an elegant and user-friendly interface. It enables users to input a news article and select from a choice of two models to generate a comprehensive list of relations within the provided text. As a result, RENA allows real-time parsing of news articles to extract key information for epidemic surveillance, contributing to EPIWATCH, an open-source intelligence-based epidemic warning system.

Introduction

Online news websites are a valuable source of information about real-world events with several potential applications. One such application is the use of news articles for text-based epidemic intelligence (Joshi et al. 2019). EPIWATCH is an open-source intelligence-based early warning system for epidemics that parses vast amounts of data in real-time to build a structured repository of data to study changing trends in diseases and syndromes, geographically, over time (MacIntyre et al. 2023). A relevance classifier selects news articles from multiple sources (**?**), which are then reviewed by expert analysts daily to generate structured data such as name of disease, number of cases and so on. This structured information allows identification of potential outbreaks earlier than traditional surveillance systems, alerting health authorities, thereby allowing for quicker outbreak response, and preventing further spread, ultimately saving lives (Puca and Trent 2020).

The manual task of extracting structured information from news articles maps to the natural language processing (NLP) task of relation extraction (Banko and Etzioni 2008; Kumar 2017). Our browser-based tool, *Relation Extraction from News Articles (RENA)*, uses decoder-only foundation models (also known as 'large language models', *i.e.*, LLMs) to extract semantic relations in infectious disease-related news articles in the English language[1] at the document level. When an epidemiologist enters a news article, RENA extracts a list of entities and relations present in the article. This streamlines an automated process for epidemiologists and researchers who need a method to acquire a large congregate of structured relations from their selected article. By making use of RENA, they can be aided in their research without having to read a large set of news articles or delve into manual data curation.

Whilst many previous relation extraction (RE) tasks have focused on the sentence level, it is evident that many relations exist between different sentences, presenting another challenge in extracting relations at the document level (Xu, Chen, and Zhao 2021). We assume a simplistic definition of a relation: a relation connects exactly two entities. In the context of epidemic intelligence, the sentence 'A patient died due to COVID-19 today' results in the relations {death number: '1', relation: 'death of', infectious disease: COVID-19'}, {death number: '1', relation: 'occurred on', event date: 'today'} and {infectious disease: COVID-19, relation: 'occurred on', event date; 'today'}. In this specific case, if two relations are identified, the third can be inferred.

RENA can be used by epidemiologists, public health officials and teachers or students of public health to extract information from news articles of interest. While the utility of RENA is for infectious disease-related epidemic intelligence, it can potentially be used for news articles across domains and application areas, including journalists/news publishers who want to verify and investigate information across multiple sources.

Architecture

Figure 1 shows the architecture of RENA. In the model training phase, we generated a set of 300 synthetic articles, annotated with relevant entities and relationships

*These authors contributed equally.

†Corresponding author

[1]With appropriate choice of foundation models, RENA can be extended to languages other than English. We are interested in doing so.
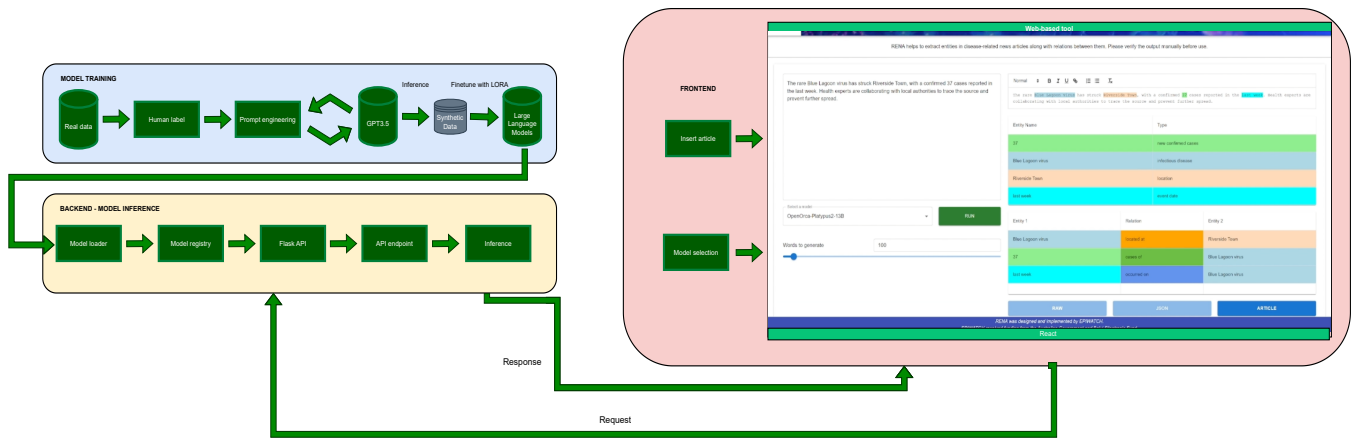
Figure 1: Architecture

using 1-shot prompting with OpenAI API [2] (gpt-3.5-turbo-16k) (Details in Supplementary Document.). The generated output (articles, and labeled entities and relations) were used for fine-tuning two large language models including `OpenOrca-Platypus2-13B`[3] and `Mythical-Destroyer-V2-L2-13B`[4], both based on Meta's LLaMa2 13B model (Touvron et al. 2023), using quantized low-rank adaptation (QLoRa) (Dettmers et al. 2023). We run the training process for three epochs. In the inference phase, RENA allows for the selection of the two different models specified above. Our backend is built with Python Flask, where the model is both loaded and executed for inference. The output is then sent to our frontend, developed in React, to display the response.

We evaluated the two models included in RENA on 10 randomly selected articles in English, selected from the EPIWATCH database, an existing database with more than 1 million news articles relating to infectious disease outbreaks searched through online news, and prioritised using machine learning-based classification supplemented with manual curation by experts. We computed precision, recall and F1 score, comparing the predicted to labeled output on the ten articles for each model. We assessed the accuracy for named entity recognition (NER) and RE separately. RE was only evaluated for relevant entities that were recognised, so both metrics should be considered when assessing the models. Table 1 shows the precision, recall and F-score values for the models included in RENA.

User Interface

Figure 1 is a snapshot of the web interface hosting RENA. On the webpage includes:

- **Input:** The input contains the text field that the user can input their unstructured article into, from which the

---

Model	Eval	Precision	Recall	F1
Open-Orca-13b	NER	0.93	0.66	0.77
Open-Orca-13b	RE	0.88	0.88	0.88
Mythical-Destroyer-13b	NER	0.96	0.57	0.71
Mythical-Destroyer-13b	RE	0.97	0.97	0.97

Table 1: Evaluation of models in RENA.

model will extract entities and relations.

- **Select Model:** The select model box contains a dropdown to select between the 2 models to be used for generation.
- **Submit:** Begins the generation process.
- **Max tokens:** Max tokens is a slider to determine how many tokens of output will be produced by the model.
- **Output:** The model will generate the output (relations) into this box.
- **Raw Button:** The raw generated output of the model will be displayed in the output box when clicked.
- **JSON Button:** The output will be converted into a JSON format and be displayed in the output box when clicked.
- **Article Button:** The article will be displayed in the output box and the entities will be highlighted in the text and color coded. A table showing all the entity and entity types as well as another table showing the relations will also be shown.

Summary & Future Work

RENA is a tool that uses QLoRA to finetune LLMs and extract disease-related relations from news articles. It offers a way to automate the creation of structured data from an inputted unstructured news article and output relatively accurate results. RENA will enable users interested in researching diseases to parse large amounts of data to potentially aid in the surveillance and prevention of diseases. We plan to extend RENA to multilingual news articles. In addition, different relations can help delineate different disease events,

allowing detection of independent outbreaks involving different infections or locations from news articles.

References

Banko, M.; and Etzioni, O. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, 28–36.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Joshi, A.; Karimi, S.; Sparks, R.; Paris, C.; and Macintyre, C. R. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6): 1–19.

Kumar, S. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*.

MacIntyre, C. R.; Chen, X.; Kunasekaran, M.; Quigley, A.; Lim, S.; Stone, H.; Paik, H.-y.; Yao, L.; Heslop, D.; Wei, W.; et al. 2023. Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, 51(3): 03000605231159335.

Puca, C.; and Trent, M. 2020. Using the Surveillance Tool EpiWATCH to Rapidly Detect Global Mumps Outbreaks. *Global Biosecurity*, 2(1).

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xu, W.; Chen, K.; and Zhao, T. 2021. Document-level relation extraction with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14167–14175.