Proposal for Text Analytics
Entity-Relation-Extraction from Text

Team Members:

Xin SUN (3528729) -- ethan.sun921107@gmail.com
Yuerui YAN (3585263) -- elinay0126@gmail.com
Yebei HU (3565290) -- yebeihu@gmail.com

GitHub :
https://github.com/XIN-von-SUN/ITA-Project.git

● Motivation and Objective

Information Extraction (IE) refers to extract structural information from natural language text, which is a promise for various NLP tasks, such as Question-Answering System, Machine Comprehension, Knowledge Graphs, Text Summarization, etc. And Named entity recognition (NER) and relation extraction (RE) are two important sub-tasks in information extraction.

Entity-Relation-Extraction More specifically, NER aims to locate and classify named entities mentioned in unstructured text into predefined categories such as organizations, person names,, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. And relationship extraction requires the identification and classification of semantic relationship mentions within a set of entities, typically from text or structural documents.

Objectives Subsequently, the extraction of named entities(NER) and their semantic relations (relation extraction, RE) are key tasks in information extraction and retrieval (IE & IR). For instance, given a textual sentence pr document, the objective is to identify both the named entities and the relations between them as we propose to do in this project.

● Available Data

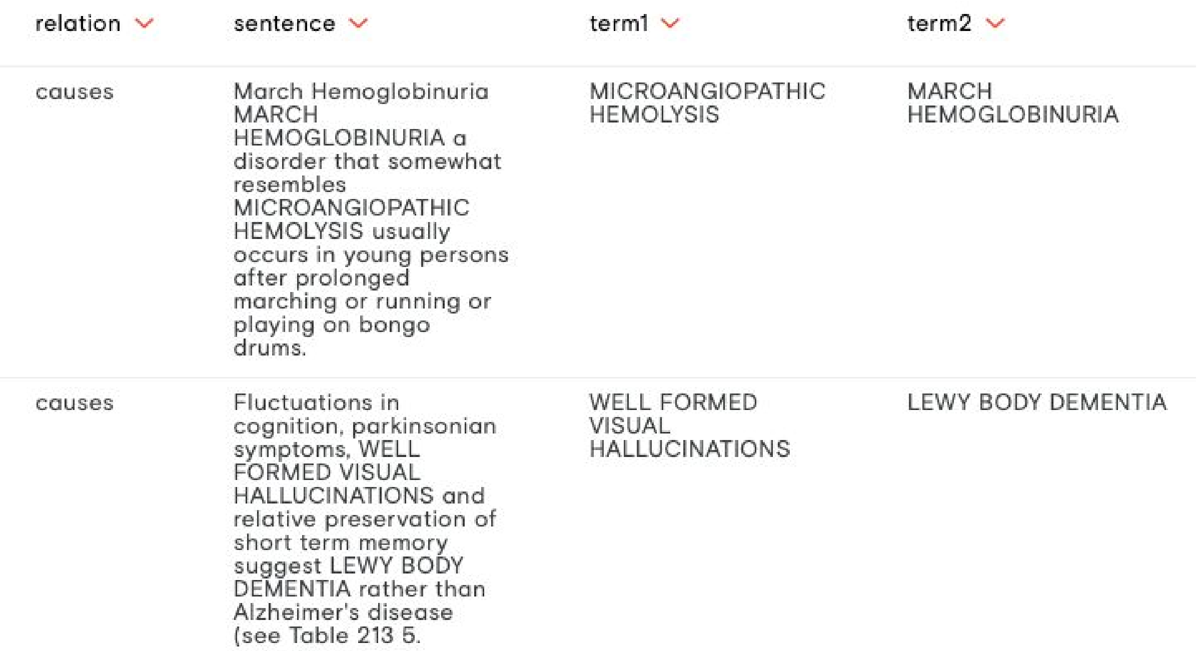
There are several free open-source dataset online for NER and relation extraction. We temporarily chose two of them as the candidate dataset for our proposed task.

Dataset 1: appen - Medical Information Extraction
This is an open-source dataset can be used for medical terminology identification and also medical relation extraction. We can do experiments on this dataset.

This dataset[1] contains 3,984 medical sentences extracted from PubMed abstracts and relationships between discrete medical terms were annotated. This dataset focuses primarily on “treat” and “cause” relationships, with 1,043 sentences containing treatment relations and 1,787 containing causal ones. We can see the example in Figure 1.

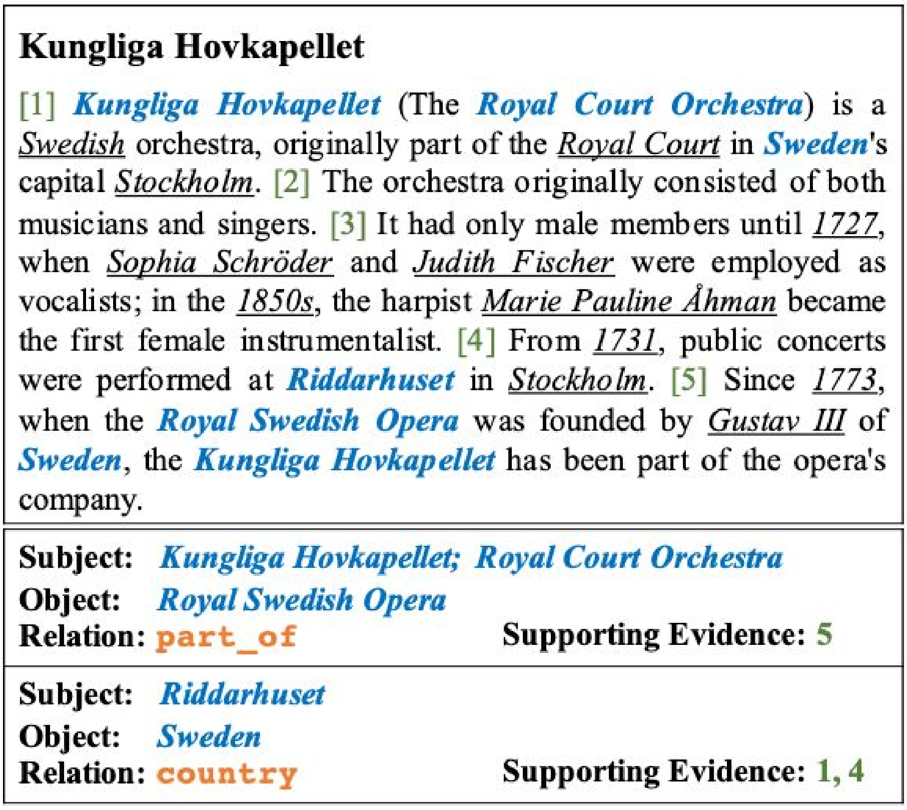
Dataset 2: DocRED - Document-level RE

DocRED[2], a new dataset constructed from Wikipedia and Wikidata with three features: (1) DocRED annotates both named entities and relations, and is the largest human- annotated dataset for document-level RE from plain text; (2) DocRED requires reading mul- tiple sentences in a document to extract entities and infer their relations by synthesiz- ing all information of the document. We can see the example in Figure 2.



relation	sentence	term1	term2
causes	March Hemoglobinuria MARCH HEMOGLOBINURIA a disorder that somewhat resembles MICROANGIOPATHIC HEMOLYSIS usually occurs in young persons after prolonged marching or running or playing on bongo drums.	MICROANGIOPATHIC HEMOLYSIS	MARCH HEMOGLOBINURIA
causes	Fluctuations in cognition, parkinsonian symptoms, WELL FORMED VISUAL HALLUCINATIONS and relative preservation of short term memory suggest LEWY BODY DEMENTIA rather than Alzheimer's disease (see Table 213 5.	WELL FORMED VISUAL HALLUCINATIONS	LEWY BODY DEMENTIA

Figure 1. Dataset 1 - appen [1]



Kungliga Hovkapellet

[1] *Kungliga Hovkapellet* (The *Royal Court Orchestra*) is a *Swedish* orchestra, originally part of the *Royal Court* in *Sweden*'s capital *Stockholm*. [2] The orchestra originally consisted of both musicians and singers. [3] It had only male members until *1727*, when *Sophia Schröder* and *Judith Fischer* were employed as vocalists; in the *1850s*, the harpist *Marie Pauline Åhman* became the first female instrumentalist. [4] From *1731*, public concerts were performed at *Riddarhuset* in *Stockholm*. [5] Since *1773*, when the *Royal Swedish Opera* was founded by *Gustav III* of *Sweden*, the *Kungliga Hovkapellet* has been part of the opera's company.

Subject: *Kungliga Hovkapellet; Royal Court Orchestra*
Object: *Royal Swedish Opera*
Relation: `part_of` **Supporting Evidence:** 5

Subject: *Riddarhuset*
Object: *Sweden*
Relation: `country` **Supporting Evidence:** 1, 4

Figure 2. Dataset 2 - DocRED [2]

Potential Approaches

Since we are not sure if we can finish the ultimate objective for Entity-Relation extraction on Text. Therefore, we divide the overall task into three phases: Entity Extraction, Relation Extraction and jointly Entity and Relation Extraction.

Phase 1. Entity Extraction

In this first phase of our project, we aim to do entity resolution and extraction on the texts. For instance in dataset 1 - appen, to extract all medicine terminologies from the sentences is our objective in this phase.

➔ Possible methods:

1. CRF+LSTM based NER
2. (Advanced Trial) Pre-trained model (BERT) for sequence labelling as NER

Phase 2. Relation Extraction

There are various different possible methods for doing Relation Extraction as below[3]:

- Rule-based Relation Extraction;
- Weakly Supervised Relation Extraction;
- Supervised Relation Extraction;
- Distantly Relation Extraction;
- Unsupervised Relation Extraction;

For our task, we propose to apply rule-based and supervised methods for experiment.
For instance, in dataset 1, we can predefine some fixed pattern for extracting the relation between given entities. Then we can also try to use DNN or pre-trained models as feature extractor for doing relations classification in a supervised way.

Phase 3 (Exploratory Task). Jointly Entity and Relation Extraction

Nowadays, the IE system can combine the NER and RE modules as an assemble model for doing extraction. Therefore, we also want to do an exploratory attempt for jointly extracting the entity and relation from text as this paper did [4].

Evaluation Metrics

For the entity and relation extraction, the system performance is usually reflected using the performance measures of information retrieval: precision, recall, and F-measure (van Rijsbergen, 1979). Precision is the ratio of the number of correctly predicted positive examples to the number predicted positive examples. Recall is the ratio of the number of correctly predicted positive examples to the number of true positive examples. F-measure (Fm) combines precision and recall as follows:

$$Fm = \frac{2 * precision * recall}{(precision + recall)}$$

- **Reference**

[1] Appen-Medical Information Extraction
(https://appen.com/datasets/medical-sentence-summary-and-relation-extraction/)

[2] DocRED: A Large-Scale Document-Level Relation Extraction Dataset
(https://arxiv.org/pdf/1906.06127.pdf)

[3] Different ways of doing Relation Extraction from Text
(https://medium.com/@andreasherman/different-ways-of-doing-relation-extraction-from-text-7362b4c3169e#:~:text=Relation%20Extraction%20(RE)%20is%20the,%2C%20is%20in%2C%20France)

[4] END-TO-END NAMED ENTITY RECOGNITION AND RELATION EXTRACTION USING PRE-TRAINED LANGUAGE MODELS
(https://arxiv.org/pdf/1912.13415.pdf)