Executive Summary

Large Language Models and Topic Modelling for Textual Information Extraction and Analysis

Marcos Tidball
GitHub: @zysymu
LinkedIn: in/zysymu

Key Findings

- "Big narratives" (a custom prompt in human-readable format that uses a combination of tabular data and narratives) provide LLMs a more complete understanding of what happened.
- Text2Text Generation transformers that combine "big narratives" with specific questions allow for the automatic extraction of information present only on the narratives, such as precipitating events, activity involved and more complete diagnoses.
- Topic modelling can be used to create categories from information extracted from narratives.

Summary of Approach

Data Sources

The only data used is the Primary Data. I "translate" the most frequent technical terms and expressions present in the narratives into common terms (Fig. 1). This step was fundamental for optimal results when using pretrained Text2Text Generation LLMs.

Methodology

Inspired by TabLLM, I create custom prompts ("big narratives") that combine patient information contained on tabular variables with the translated narrative (Fig. 2). I combine each big narrative with 6 questions (Fig. 3) and use the `google/flan-t5-base` model for Text2Text Generation; this allows the LLM to extract specific information about the falls based on the questions.

To analyze the answers, I trained one Latent Dirichlet Allocation (LDA) model per question to model the topics in the answers. This allows answers to be categorized into topics with similar subjects, which, together with the 6 different question categories provide a very granular way to analyze the data. I preprocessed the answers with: word-level tokenization, stop words removal, creation of bigrams and trigrams and lemmatization. However, even after hyperparameter optimization, some topics are still not very clear. This is probably due to the loss of positional and textual information caused by the preprocessing pipeline. I believe that using a more robust topic modelling algorithm such as BERTopic would provide better results.

Evaluation

For the information extraction part, I tried different Question Answering models, but due to their extractive nature (answers could only be text present in the prompt), sparse narratives were problematic. I also tried using just the narratives available on the base dataset with and without the translation step, but the translated big narratives allowed for more complete and accurate answers, at least in a sample of 50 randomly selected narratives that I manually analyzed.

The LDA models were tuned only for coherence in mind, with their performance being evaluated via the coherence score, visualizations and by checking the most representative answers for each topic (Fig. 4). Initial analyses and plots that demonstrate how these topics can be used are shown in Fig. 5 and Fig. 6.

Visualizations

```
medical_terms = {
    "@": "at",
    ">>": "clinical diagnosis",
    "&": "and",
    "***": "",
    "dx": "clinical diagnosis",
    "fx": "fracture",
    "pt": "patient",
    "c/o": "complains of",
    "lt": "left",
    "rt": "right",
    "l": "left",
    "r": "right",
    "loc": "loss of consciousness",
    "s/p": "after",
    "inj": "injury",
    "nh": "nursing home",
    "w/o": "without",
    "w/": "with",
    "eval": "evaluation",
    "sts": "sit-to-stand",
    "bal": "blood alcohol level,",
    "ecf": "extended-care facility",
    "bac": "blood alcohol content",
    "etoh": "ethyl alcohol",
    "clsd": "closed",
    "afib": "atrial fibrillation",
    "er": "emergency room",
    "hx": "history of",
    "uti": "urinary tract infection",
    "pta": "prior to arrival",
    "biba": "brought in by ambulance",
    "glf": "ground level fall",
    "inr": "international normalized ratio",
    "h/o": "history of",
    "mech": "mechanical",
    "ams": "altered mental status",
    "mult": "multiple",
    "fxs": "fractures",
    "n.h.": "nursing home",
    "intox": "intoxication",
    "sdh": "acute subdural hematoma",
    "r/o": "rules out",
    "p/w": "presents with",
    "pts": "patient's",
    "cva": "cerebral vascular accident",
    "cpk": "creatine phosphokinase",
    "abd": "abdomen",
    "fd": "fall detected",
    "htn": "hypertension",
    "sah": "subarachnoid hemorrhage",
    "af": "atrial fibrillation",
    "px": "procedure",
    "s'd&f": "slipped and fell",
    "chi": "closed-head injury",
    "t'd&f": "tripped and fell",
    "af": "accidental fall",
    "tr": "trauma",
    "aki": "acute kidney injury",
}
```

Figure 1: Translation dictionary for technical terms.

```
The patient has suffered a fall. The patient is a 81 years old male.
The things involved in the accident were: beds or bedframes, other or not specified, , .
The patient reports fracture on its upper leg.
This is the description of the incident: patient fell out of bed. clinical diagnosis: left femur fracture.
```

Figure 2: Example of automated big narrative generated for `cpsc_case_number` = 230217170. Original narrative is: "81 YOM FELL OUT OF BED. DX: LEFT FEMUR FRACTURE."

```
questions = {
    "action": "What was the patient doing at the time of the incident?",
    "cause": "What caused the patient to fall?",
    "diagnosis": "What is the full diagnosis?",
    "how": "How did the fall occur?",
    "what": "What happened to the patient during the incident?",
    "where": "Where did the fall occur?",
}
```

Figure 3: Questions that, combined with the big narratives, form the prompts fed to the LLM.

```
'patient    last night patient slipped on wet wood surface of a  deck, fell landing on left side. clinical diagnosis: closed fracture of hip    #',
'patientcomplains of head and back injury. she walked in the front door and her foot slipped out from under her from having wet shoes and she fell onto her back.
sustained laceration to scalp. clinical diagnosis: fall, head injury, scalp laceration',
'patient from the nursing home was walking with her walker when her foot caught under the wheel of the walker and she fell to the floor onto her head. clinical d
iagnosis: closed head injury',
'patient was sitting in her chair and when she went to get up she got her foot caught on the wheel of her walker causing her to fall. clinical diagnosis: closed
nondisplaced fracture of fifth metatarsal bone of left foot and closed nondisplaced fracture of proximal phalanx of left great toe',
'patient,  at  extended-care facility, ambulating with her walker when her foot got caught on the back wheel of the walker and fell to hard floor. clinical diagn
osis: wrist fracture'
```

Figure 4: Most representative translated narratives (highest score) for topic 0 of the `cause` question. The keywords for this topic are: "slip, rugs_carpet, ceilings_wall, bathtubs_shower, frame, door_sill, shoe, walker, foot, catch".

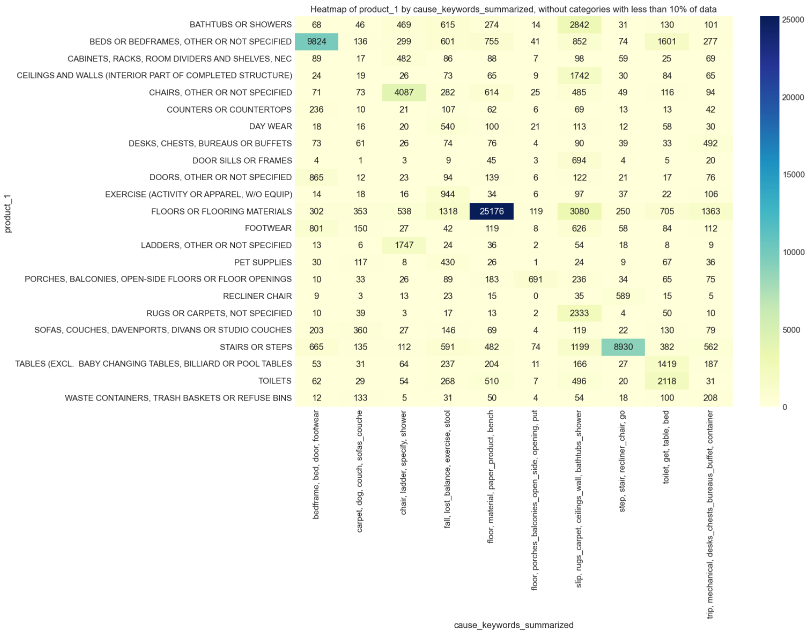

Figure 5: Heatmap of `product_1` by `cause` topics (with each topic's keywords being displayed for better explainability) displaying only categories with more than 10% of total data for better readability. We can notice that there is a good match between the `product_1` categorical variable and the more specific `cause` topics, which give us more information about what happened.

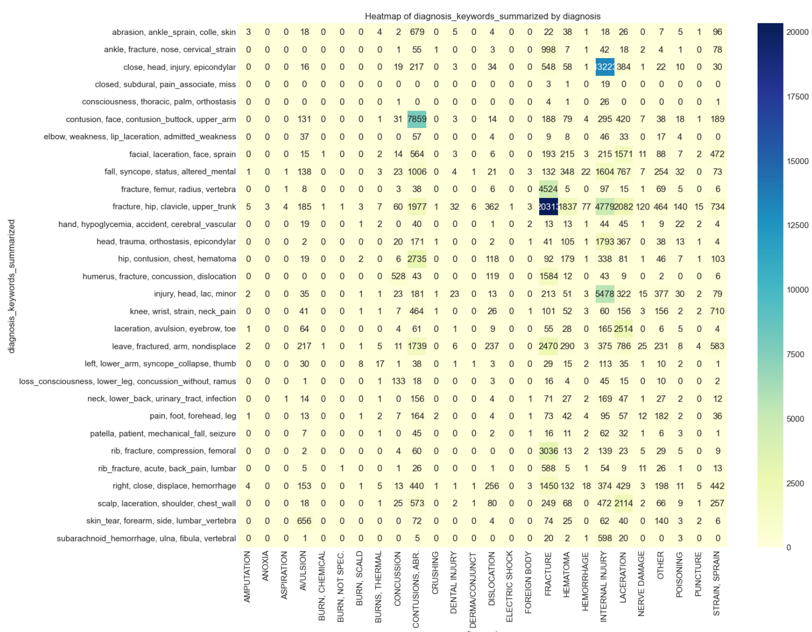

Figure 6: Heatmap of `diagnosis` by `diagnosis` topics (with each topic's keywords being displayed for better explainability). Visualizations like this can be used to evaluate if the topics make sense. This heatmap demonstrates that the information extraction pipeline that was developed can aid in obtaining more specific details about what happened in the fall events.