Assignment 3: Knowledge Graph Population

Entity Extraction for Hospital Resource Management

Course: Knowledge Graphs with Large Language Models **Program**: MSc in AI and Data Science, 2025-2026 **Instructor**: Panos Alexopoulos

---

1. Introduction

This assignment implements an LLM-based entity extraction system for populating a Hospital Resource Management knowledge graph. We extract two entity types:

- **Equipment**: Medical devices and equipment (MRI machines, CT scanners, ventilators)
- **Department**: Hospital departments (Emergency Department, Cardiology, ICU)

The system uses GPT-4o with few-shot prompting (no fine-tuning required).

---

2. Methodology

2.1 Entity Extractor (Task 1)

Approach: Few-shot prompting with GPT-4o

The extractor receives a prompt containing:

- Entity type definitions
- Extraction rules (only extract what appears in text)
- 3 example extractions
- JSON output format

Configuration:

- Model: GPT-4o
- Temperature: 0.0 (deterministic)
- Output: JSON with Equipment and Department lists

2.2 Evaluation Dataset (Task 2)

Dataset:

- 12 texts from real hospital sources
- 7 from actual press releases (2024)
- 5 synthesized realistic examples
- Sources: UCI Health, Northwestern Medical Center, UNM Hospital, St. Peter's Hospital, etc.

Annotation:

- Manual ground truth annotation

- Total: 71 Equipment entities, 48 Department entities
- Entities extracted as they appear in text

Metrics:

- Precision: Fraction of extracted entities that are correct
- Recall: Fraction of ground truth entities found
- F1 Score: Harmonic mean of precision and recall

2.3 LLM-as-a-Judge (Task 3)

An LLM evaluator that judges if extracted entities are:

- CORRECT: Appears in text and correctly classified
- INCORRECT: Hallucinated or wrong type
- PARTIAL: Partially correct but vague

---

3. Results

3.1 Overall Performance

Metric	Score
Precision	0.745
Recall	0.767
F1 Score	0.750

The system achieved 75% F1 score, indicating good balanced performance.

3.2 By Entity Type

Equipment:

- Precision: 0.897
- Recall: 0.903
- F1: 0.899

Department:

- Precision: 0.593
- Recall: 0.631
- F1: 0.601

Equipment extraction (90% F1) significantly outperformed department extraction (60% F1).

3.3 Performance Breakdown

Best texts (F1 = 1.0): Texts 2, 6, 8, 9, 12

- Clean, explicit mentions of equipment and departments

- Standard terminology

Challenging texts: Texts 4, 5

- Confusion between specialty names ("gynecology", "urology") and department names
- The extractor did not classify specialty types as departments

Example - Text 8 (Perfect F1 = 1.0):

- Input: "*Emergency Department upgraded with portable ultrasound devices and ECG monitoring system. Radiology collaborated with Emergency. Cardiology received echocardiography machines.*"
- Extracted: 6/6 equipment correct, 5/5 departments correct

3.4 LLM Judge Results

The judge successfully identified:

- Correct extractions matching the text
- Over-generalizations (e.g., "new equipment" too vague)
- No hallucinations detected in our extractions

---

4. Discussion

What Worked

1. **Equipment extraction (90% F1)**: Strong performance, few-shot examples effectively taught the model
2. **No hallucinations**: System only extracted terms present in text
3. **Standard terms**: Common equipment (CT scanner, MRI, ventilators) and departments (ICU, Emergency) extracted reliably

Challenges

1. **Specialty vs. Department confusion**: Main issue distinguishing between:
   - Department names: "Surgery Department", "Cardiology"
   - Specialty types: "gynecology", "thoracic surgery"
2. **Department name variations**: Text mentioning "interventional radiology suites" - unclear if referring to department, facility, or subspecialty
3. **Annotation inconsistency**: Some texts annotated specialties as departments, others didn't

Limitations

- Small dataset (12 texts)
- Limited to two entity types
- No entity relationship extraction
- No synonym handling

---

5. Conclusions

The LLM-based entity extractor achieved 75% F1 score overall, with particularly strong performance on equipment extraction (90%). Few-shot prompting with GPT-4o proved effective without requiring fine-tuning.

The main challenge was semantic ambiguity (specialty names vs. department names) rather than technical extraction capability. This could be addressed with:

- Clearer annotation guidelines
- Two-stage extraction (extract then classify)
- Larger evaluation dataset

The system demonstrates that prompt engineering is sufficient for practical knowledge graph population tasks.
