Domain-Specific Named Entity Recognition

Garvika Singh Rajawat, Aditi Singh, Kruti Patel and Hani Soni

Abstract

Domain-specific Named Entity Recognition (NER) is critical for extracting meaningful entities from unstructured text. Existing models often fall short in specialized fields due to the lack of domain-specific annotations. This project addresses these challenges by developing and training an NER model for the Medical and Solar Panel Technology Patent domains. We address data sparsity by employing GLiNER, a lightweight, zero-shot learning-based NER system pre-trained on synthetic data, to initiate annotations on an unlabeled corpus. Active learning techniques, including least confidence sampling and iterative manual validation, were used to expand the labeled dataset effectively. Model training and fine-tuning were conducted using SpaCy's pipeline, with custom labels and configurations tailored to the target domains. Our results demonstrate the efficacy of leveraging synthetic data, active learning, and domainspecific fine-tuning for NER in resource-constrained environments. This approach improves entity extraction accuracy and establishes a scalable workflow for other domain-specific NER tasks.

Keywords

Named Entity Recognition, Natural Language Processing, Machine Learning, Text Mining, Information Retrieval.

1. Introduction

The ever-growing volume of domain-specific data has created a pressing need for efficient tools to extract meaningful entities from unstructured text. Named Entity Recognition (NER), a crucial Natural Language Processing (NLP) task, identifies and categorizes entities within text. However, existing NER models trained on general datasets often fail to capture domain-specific terminologies, resulting in subpar performance in specialized fields. This project focuses on developing a robust NER model for two distinct domains: Medical and Solar Panel Technology Patents. For the Medical domain, entities such as Chemicals, Phenotypes, and Anatomical Structures are critical for biomedical research, drug discovery, and clinical data analysis. In the Solar Panel Technology domain, entities like Components, Manufacturing Processes, and Performance Metrics are essential for analyzing technological advancements and patent trends.

2. Methodology

The proposed methodology is structured as follows:

2.1. Problem Definition

- Extract entities from unstructured text in the Medical and Solar Panel Technology domains.
- Address challenges, including a lack of labeled data, high annotation costs, and domain expertise requirements.

2.2. Data Annotation

- **GLiNER Zero-Shot:** Initial annotation of 200 samples using pre-trained GLiNER.
- **Manual Validation:** Verified 200 samples for domain-specific accuracy.
- **Active Learning:**

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318061@daiict.ac.in (G. S. Rajawat); 202318005@daiict.ac.in (A. Singh); 202318032@daiict.ac.in (K. P.); 202318044@daiict.ac.in (H. Soni)
202318061 (G. S. Rajawat); 202318005 (A. Singh); 202318032 (K. P.); 202318044 (H. Soni)

  - Trained the model on validated data.
  - Identified 200 least confident sentences for manual validation.
  - Iteratively trained with new samples added to previous batches.

2.3. Model Training with SpaCy

- **Data Preparation:** Labeled data was converted to SpaCy's training format.
- **Pipeline Configuration:** Customized for domain-specific labels.
- **Fine-Tuning:** Trained using SpaCy's NER pipeline.
- **Evaluation:** Metrics included Precision, Recall, and F1-Score.
- **Inference:** Deployed on unlabeled corpora for entity extraction.

2.4. Tools Used

- **GLiNER:** Lightweight zero-shot learning for initial annotation.
- **SpaCy:** Comprehensive NLP pipeline for training, fine-tuning, and inference.

3. Results and Evaluation

This section presents the evaluation results for the Named Entity Recognition (NER) models trained on Medical and Solar Panel Technology data. The performance metrics, including Precision, Recall, and F1-Score, are detailed below.

3.1. Medical Data Evaluation

- **Batch:** 7
- **Metrics:**

```
"overall": {
    "precision": 0.7537,
    "recall": 0.6044,
    "f1": 0.6727,
    "sup": 2117,
    "predicted": 1731
}
```

3.2. Solar Data Evaluation

- **Batch:** 4
- **Metrics:**

```
"overall": {
    "precision": 0.5137046861184792,
    "recall": 0.27444496929617385,
    "f1": 0.3577586206896552,
    "sup": 1310,
    "predicted": 1131
}
```

4. Conclusion

This study demonstrates the application of advanced NER techniques to achieve improved entity recognition accuracy in resource-constrained environments. The results highlight the importance of feature selection and domain adaptation for real-world text datasets. Future work includes scaling this approach to other specialized domains and further optimizing active learning workflows.

References

- Jinhyuk Lee, et al. *BioBERT: A pre-trained biomedical language representation model for biomedical text mining*. Bioinformatics, 36(4):1234–1240, 2020.
- Urchade Zaratiana, et al. *GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer*. arXiv:2311.08526 [cs.CL]. https://doi.org/10.48550/arXiv2311.08526, 2023.
- Andrew Eliot Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. PhD Thesis, New York University, 1999. AAI9945252.