Multilingual and Synthetic Medical NER for Low-Resource Languages

Neural Networks - Intermediate Presentation

Matei Calin Popescu (matei_calin.popescu@stud.acs.upb.ro)
Alexandru Rus (alexandru.rus0704@stud.acs.upb.ro)

What is NER and Why Does It Matter?

Named Entity Recognition (NER) detects and classifies important entities in text:

- People, places, organizations
- Domain-specific terms (medical conditions, treatments, etc.)

Our Project Journey

Started with: Romanian NER using the RONEC dataset

Expanded to include:

- Multilingual approach (Romanian + French)
- Medical domain focus
- Synthetic data generation using ChatGPT

> **Research Question:** Can synthetic medical data and multilingual transfer learning help improve NER for languages with limited annotated resources?

Background & Related Research

BERT-based models are now the standard for NER tasks

- Multilingual BERT (mBERT) learns shared representations across languages
- Works well even for languages with small training datasets

Low-resource languages benefit from:

- Cross-lingual transfer (training on multiple languages together)
- Synthetic data to supplement limited real examples

Medical NER is challenging:

- Most research uses English datasets (i2b2, MEDLINE)
- No large annotated medical corpora exist for Romanian or French
- This makes synthetic generation especially valuable

Datasets We Used

1. RONEC (Romanian Named Entity Corpus)

- Real, manually annotated data
- General domain text (news, etc.)
- Entity types: Person, Organization, Location

2. Synthetic Romanian Medical Data

We generated medical sentences using ChatGPT with realistic clinical scenarios

> **Example:**
> "Dr. Ionescu tratează pacienți cu insuficiență cardiacă la Spitalul Elias."
>
> **Dr. Ionescu** → Person
> **Spitalul Elias** → Organization
> **insuficiență cardiacă** → Medical condition

Datasets (continued)

3. Synthetic French Medical Data

Generated similar medical sentences in French to test cross-lingual transfer

> **Exemple 1:**
> "Le Dr. Martin travaille à l'Hôpital Saint-Louis et traite des patients atteints de diabète."
>
> **Dr. Martin** → Person **Hôpital Saint-Louis** → Organization **diabète** → Medical

> **Exemple 2:**
> "La patiente a été transférée à la Clinique Pasteur pour un examen spécialisé."
>
> **Clinique Pasteur** → Organization

These synthetic datasets let us test whether adding French medical examples improves Romanian medical NER

How We Set Up the Experiments

Model & Training

Model: bert-base-multilingual-cased

- Token classification approach
- BIO tagging scheme

Training parameters:

- Learning rate: 5e-5
- Batch size: 16
- 3 epochs
- AdamW optimizer

Three Experiments

Experiment 1: Baseline
Train only on RONEC

Experiment 2: Add Romanian Medical
RONEC + synthetic Romanian medical data

Experiment 3: Add French Too
RONEC + Romanian medical + French medical

Goal: see if synthetic data helps, and whether multilingual training improves things further

Results So Far

Experiment	Precision	Recall	F1
Baseline (RONEC only)	0.81	0.78	0.79
+ Synthetic Romanian medical	0.83	0.82	**0.825**
+ Romanian + French medical	0.85	0.84	**0.845**

Key takeaways:

- Adding synthetic Romanian medical data: **+3.5 points** F1
- Adding French on top of that: **+5.5 points** total F1 improvement
- Both synthetic data and multilingual transfer appear helpful

What's Next

What we've done so far:

- Expanded from Romanian-only to multilingual (RO + FR)
- Created synthetic medical datasets with ChatGPT
- Ran initial experiments showing promising improvements

Still need to:

- Run full training with more epochs and proper validation curves
- Do ablation studies (what happens with ONLY Romanian synthetic? ONLY French?)
- Look at where the model makes mistakes (error analysis)
- Test on real medical text if we can find some
- Write up final report and polish the presentation

These initial results suggest that synthetic data generation and multilingual training could be a viable path for improving medical NER in low-resource settings

Questions?

Matei Calin Popescu - matei_calin.popescu@stud.acs.upb.ro
Alexandru Rus - alexandru.rus0704@stud.acs.upb.ro