Day 12: Named Entity Recognition (NER) for Urdu & Pashto

Sami Uddin Shinwari
Subject Specialist – IT | NLP Researcher
MS Data Science (FAST NU Peshawar)
BS Computer Science (UET Peshawar)

January 11, 2026

What is Named Entity Recognition (NER)?

- Identifying and classifying named entities in text.
- Common entity types:
    - Person (PER)
    - Location (LOC)
    - Organization (ORG)
- A key task in Information Extraction.

Why NER is Important

- Helps extract structured information from text.
- Improves search and information retrieval.
- Essential for knowledge graphs.
- Widely used in real-world NLP systems.

NER Example (English)

Sentence:
Imran Khan visited Islamabad

Named Entities:

- Imran Khan – Person
- Islamabad – Location

NER Example (Urdu)

Sentence:

عمران خان اسلام آباد گئے

Named Entities:

- خان عمران – (Person)
- آباد اسلام – (Location)

NER Example (Pashto)

Sentence:
عمران خان اسلام آباد ته لاړ

Named Entities:

- عمران خان – (Person)
- اسلام آباد – (Location)

Challenges in Urdu & Pashto NER

- No capitalization cues.
- Ambiguity between common nouns and names.
- Limited annotated NER datasets.
- Spelling variations in names.

Existing Tools & Resources

- CRULP Urdu NER corpus
- IJCNLP Urdu NER datasets
- Universal Dependencies (partial)
- Rule-based + ML approaches

Use Cases for Urdu & Pashto NER

- News article analysis
- Social media monitoring
- Information extraction from government records
- Question answering systems

Why NER is Hard for Low-Resource Languages

- Lack of labeled training data.
- Few pretrained language models.
- Domain and dialect diversity.
- Limited linguistic tools.

Key Takeaways

- NER turns unstructured text into structured data.
- Urdu & Pashto need more open NER datasets.
- Classical models still provide strong baselines.
- Community contribution is crucial.

#Day12 #NER #UrduNLP #PashtoNLP #LowResourceNLP