ACADEMIC YEAR 2026/27

Text Mining & Search [DS] / Information Retrieval [CS]

Named Entity Recognition

TEXT MINING & SEARCH

Prof. Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems and Communication

INFORMATION RETRIEVAL

Dr. Georgios Peikos

University of Milano-Bicocca

Department of Informatics, Systems and Communication

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Ikr3 Lab — information and knowledge representation, retrieval, and reasoning LAB

Questions?

Comments?

Named Entity Recognition

Introduction

Named Entity Recognition (NER) – What is it?

- **NER** locates entity mentions in text and assigns predefined types (categories/tags).
- A **span** is a continuous sequence of tokens or characters.
- A **mention** is an occurrence of an entity in a document.
- The **annotation schema** defines which mentions exist.

Anna Rossi joined IKR3 Labs in Milan.

Anna Rossi = PER IKR3 Labs = ORG Milan = LOC

Named Entity Recognition (NER) – Applications

Search Filtering

Search can **filter documents** by person or organization.

News Archiving

News archives can **index articles** by mentioned **entities**.

Domain Extraction

Domain systems can **extract diseases**, **drugs**, or **legal** references.

Privacy Workflows

Privacy workflows can **flag personal identifiers** for review.

Named Entity Recognition (NER) – What is its relationship with POS tagging?

Token	POS	NER
Anna	PROPN	B-PER
Rossi	PROPN	I-PER
joined	VERB	O
Acme	PROPN	B-ORG
Labs	PROPN	I-ORG

- Both tasks assign labels to positions in a token sequence.
- POS labels describe grammatical roles.
- NER labels describe mention boundaries (where it appears?How long sequence) and semantic types (person? location? etc.).

Named Entity Recognition (NER) – What entities do we have?

- PER identifies people.
- ORG identifies organizations.
- LOC identifies locations.

Other schemas distinguish geopolitical entities, products, dates, or domain concepts.

Schema choice determines the expected entity tags for tokens.

Named Entity Recognition (NER) – What entities do we have? Domain Concepts

The patient is a 57-year-old man with abdominal pain and vomiting. The pain started gradually about 20 hours ago in the epigastric and periumbilical regions, radiating to his back. He drinks around 60 units of alcohol per week and smokes 22 cigarettes per day. He is healthy with no history of allergies or using any medications. His family history is positive for type 2 diabetes (his father and sister). He lives alone and has no children. The abdomen is tender and soft. His bowel sounds are normal. His heart rate is 115/min and blood pressure 110/75 mmHg. The lab results are remarkable for leukocytosis (19.5), urea of 8.5, high CRP (145), high amylase (1200) and Glc level of 15. Cross-sectional imaging was negative for obstructive pancreatitis.

Given specific domains, we need to extract and annotate different entities (e.g. Drugs, Dosages, etc.)

Named Entity Recognition

Connection to IR

An Example of using NER in IR Systems – How is NER relevant in IR?

- **Information Retrieval (IR)** identifies information items that can satisfy a user's information need.
- The user expresses this need through a **query**.
- The query is a textual representation of the need and may express it incompletely.

The IR system **ranks** information items according to their estimated **relevance**.

Named Entity Recognition (NER) – These can be queries in Clinical IR systems

The patient is a 57-year-old man with abdominal pain and vomiting. The pain started gradually about 20 hours ago in the epigastric and periumbilical regions, radiating to his back. He drinks around 60 units of alcohol per week and smokes 22 cigarettes per day. He is healthy with no history of allergies or using any medications. His family history is positive for type 2 diabetes (his father and sister). He lives alone and has no children. The abdomen is tender and soft. His bowel sounds are normal. His heart rate is 115/min and blood pressure 110/75 mmHg. The lab results are remarkable for leukocytosis (19.5), urea of 8.5, high CRP (145), high amylase (1200) and Glc level of 15. Cross-sectional imaging was negative for obstructive pancreatitis.

Is a long description of a patient's medical history that a doctor has to use to find medical research that will help make a diagnosis and propose treatments.

Clinical Information Retrieval – Patient Descriptions as Queries

01

Clinical Need

A clinician may need **relevant medical literature**.

02

Input Format

The input can be a **long clinical note** rather than a short keyword query.

03

Rich Information

Clinical notes contain **symptoms**, **diagnoses**, **treatments**, **tests**, and patient/family **history**.

04

Context & Nuance

They also contain **negated conditions** and information about **family members**.

05

System Challenge

The system must identify which information supports the retrieval task.

Clinical Information Retrieval – The role of NER

01

Medical Entities

NER identifies medical entity mentions in the clinical note.

02

Focused Terms

These mentions can provide terms for a more focused retrieval query.

03

Context Filtering

Determines if a condition is present, negated, historical, or family-related.

04

Query Representation

The system combines selected information into a unified query representation.

05

Ranking Goal

Aims to improve ranking quality, retrieve in top positions medical papers related to the patient's disease.

Our study – Investigating query representations

01

Research Scope

Investigated how **different representations of clinical notes affect medical retrieval.**

02

Retrieval Tasks

Studied two key tasks: **clinical trial retrieval** and **medical literature retrieval**.

03

Note vs. Query

Compared complete clinical notes with queries built from **extracted information**.

04

Variables Examined

Examined **entity selection**, **negation**, **family history**, **patient history**, and **query expansion**.

05

Main Contribution

Evaluated how these choices directly affect **downstream retrieval effectiveness**.

In short, using NER, we created many queries and used them to retrieve medical articles, and we saw that there is differences in performance.

Questions?

Comments?

Named Entity Recognition

Where does an entity starts/ends?

Named Entity Recognition (NER) – Mention Boundaries

Annotated Sentence & Boundary Spans

[Anna Rossi] PER joined [Acme Labs] ORG in [Milan] LOC.

Person Mention

Anna Rossi is identified as a single person entity span.

Notice that it is 2 tokens

Organization Mention

Acme Labs is recognized as one organization mention span.

Although token labs is a normal word.

Boundary Errors

A boundary error **alters** the extracted entity span and meaning.

Character Offsets

Character offsets link extracted mentions back to the raw source text.

- Offsets represent the exact starting and ending character positions of a text span (or entity) within a larger string.

Named Entity Recognition

How can we know which tokens refer to the same entity?

Named Entity Recognition (NER) – The BIO encoding (BIO2 convention)

Token	Anna	Rossi	joined	Acme	Labs	in	Milan	.
BIO	B-PER	I-PER	O	B-ORG	I-ORG	O	B-LOC	O

- B-X begins a mention of type X. So, the first token of a named entity is B- (Begin).
- I-X continues the same mention of type X. So, any continuation token belongs to the same entity is marked as I- (Inside).
- O marks a token outside the annotated mentions. So, any token that does not belong to any named entity, is marked O (Outside).
- A one-token mention receives B-X.

Named Entity Recognition (NER) – BIO sequence constraints

Continuation Rule
I-ORG can follow B-ORG or I-ORG.

Start Condition
I-ORG cannot begin a sequence in BIO2.

Adjacent Mentions
A new adjacent mention begins with B-X, even for the same type.

Interpretability
Constraints make the extracted spans interpretable.

VALID SEQUENCE EXAMPLE

Anna	met	Luca
B-PER	O	B-PER

Correct transition: Each entity mention starts explicitly with B-PER.

INVALID SEQUENCE (CONSTRAINT VIOLATION)

Acme
I-ORG

Invalid start condition:
A sequence cannot begin with an I- tag without a preceding B-tag.

Task formulation – A quick BIO check

INPUT SENTENCE & TOKENIZATION

Anna	and	Luca	visited	New	York	.
B-PER	O	B-PER	O	?	?	O

Rules: Target entity types are **PER** and **LOC**. Punctuation "." is treated as a separate token.

Question 1

Which labels belong to "New" and "York"?

New: B-LOC (Begin Location)

York: I-LOC (Inside Location)

Together "New York" forms a multi-token Location entity mention.

Question 2

Why does "Luca" receive B-PER?

Luca starts a **new entity mention** following the conjunction "and".

Under BIO rules, every distinct named entity starts with a **B-** prefix, preventing it from incorrectly merging with "Anna".

Challenges – Context-dependent names

"Jordan thanked the audience."

Likely PER

Identified as a Person entity based on syntactic role and surrounding context.

"The team travelled to Jordan."

Likely LOC

Same surface word resolves to Location when following a preposition of destination.

"Apple opened an office."

Likely ORG

Capitalized brand name acting as subject functions as an Organization.

"She ate an apple."

Outside our schema

Common noun usage representing fruit, not a target named entity.

NER Challenges – Boundaries and nested mentions

STRUCTURED EXAMPLE: NESTED ENTITY SPANS



Outer Span (ORG): "University of Milan"

Inner Span (LOC): "Milan"

Representation: [University of [Milan]LOC]ORG

BIO TAGGING LIMITATION

Overlapping Span Conflict

1. Organization Span: "University of Milan" can form one ORG mention.

2. Location Span: "Milan" can also form a LOC mention inside that organization name.

Core Issue: A single BIO label per token cannot encode both overlapping spans simultaneously.

TASK FORMULATION

Flat vs. Nested Annotation

Flat Annotation: Restricts token assignments to a single non-overlapping boundary layer (e.g., extracting ORG only).

Nested Annotation: Requires multi-layered or span-based architectures to recognize internal entity mentions.

Takeaway: Flat annotation and nested annotation define fundamentally different tasks.

Challenges – Variation across domains and languages

Capitalization Cues

Incomplete Surface Signal

Capitalization helps in some texts but is an incomplete cue for entity extraction.

Vocabulary Limits

Out-of-Vocabulary Names

Abbreviations and unseen names challenge dictionaries.

Domain Specificity

Specialized Terminology

Medical and legal text require specialized labels, ontologies, and expert knowledge.

Domain Shift

Generalization Degradation

A model trained on news text may significantly lose effectiveness on non-news text.

Challenges – Ambiguity and annotation policy

STRUCTURED EXAMPLE: METONYMIC USAGE

The White House
LOC vs. ORG ?

issued a statement.

Context Sentence: "The White House issued a statement."

Linguistic Ambiguity: A place name can refer to an institution or government.

ANNOTATION GUIDELINES

Policy Definition

Mandatory Conventions: Guidelines must define how annotators handle this usage.

Clear policy documentation ensures consistent entity boundary and type decisions across annotators.

ERROR ANALYSIS

Disagreement Source

Evaluation Impact: Disagreement may reflect unclear policy rather than model failure.

Inter-annotator variance or evaluation penalties often stem from ambiguous specifications rather than algorithmic errors.

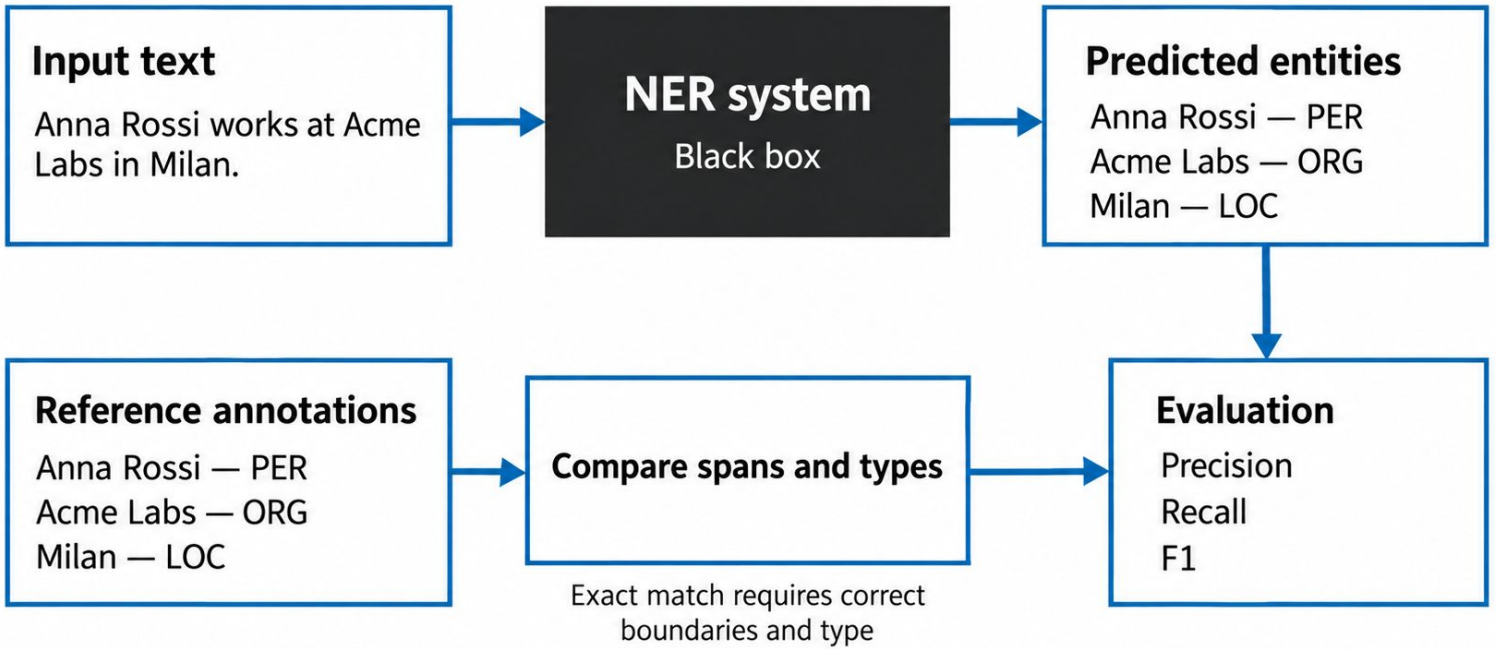
Questions?

Comments?

NER – Methods we can use to perform NER

- Rules and dictionaries identify entities using predefined patterns and lists of known names, such as cities or organizations.
- Statistical models, such as HMMs, learn from annotated examples to assign entity labels using word information and dependencies between neighboring labels.
- Neural models, such as BiLSTMs and Transformers, learn contextual representations and predict entity labels for each token.
- Generative language models, such as LLMs, follow instructions to extract entities and return their text and types in a structured format, such as JSON.

NER Systems – Overview of how such a system works



NER Methods – Rules and dictionaries

DICTIONARY LOOKUP

Known Name Lists

A dictionary lists known names and their possible entity types.

EXAMPLE MATCH

"Milan" → Location / Organization

PATTERN MATCHING

Rule-Based Clues

Patterns use clues such as capitalization or organization suffixes.

EXAMPLE PATTERN

Names ending in "Ltd" → Organization

KEY ADVANTAGE

Domain Inspection

Easy to Inspect: Rules are simple to audit, understand, and refine in a restricted domain.

KEY CHALLENGE

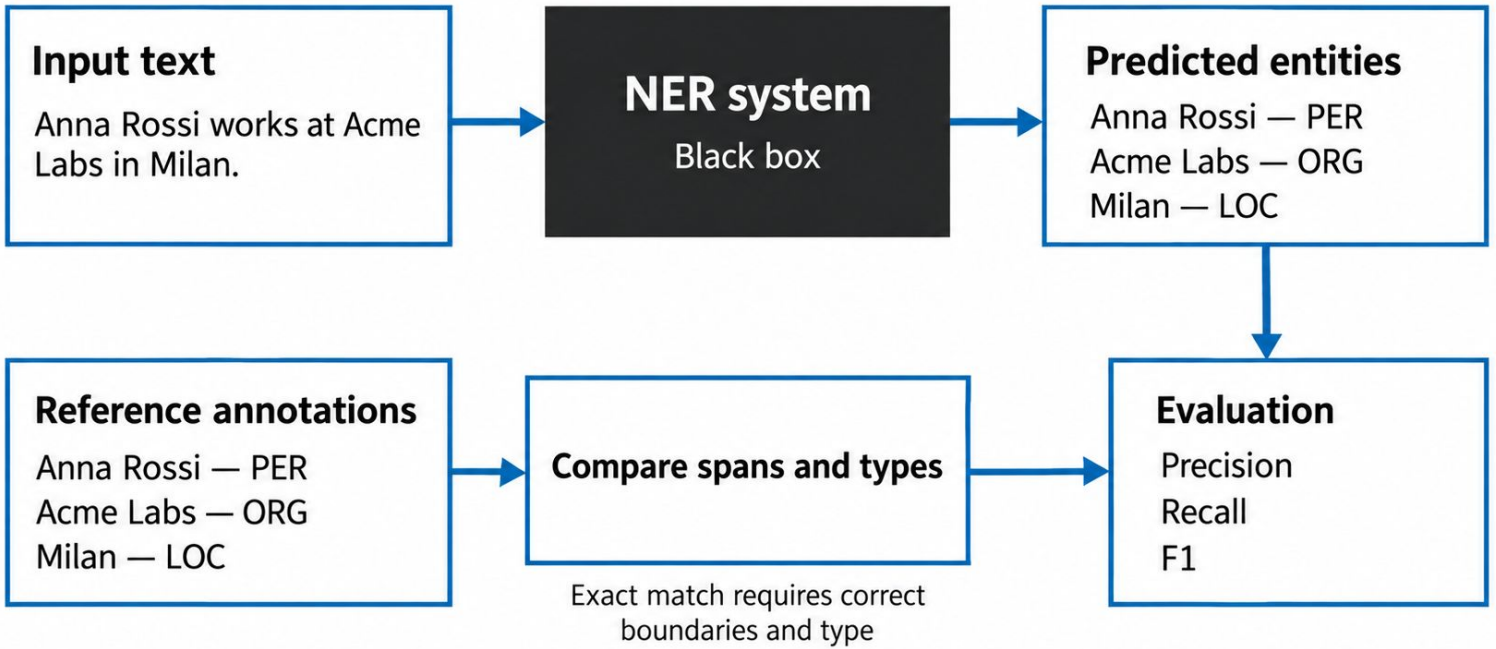
Maintenance Cost

High Maintenance: New names, ambiguous entries, and changing language require ongoing manual updates.

Questions?

Comments?

NER Systems – Overview of how such a system works



Evaluation – Exact entity matching

MATCHING ENTITY

True Positive (TP)

A true positive has the correct span and entity type.

UNMATCHED PREDICTION

False Positive (FP)

An unmatched predicted mention is a false positive.

UNMATCHED REFERENCE

False Negative (FN)

An unmatched reference mention is a false negative.

COMPOUND ERROR

Boundary & Type Errors

A boundary or type error usually contributes both FP and FN.

Evaluation – Precision, recall, and F1

FORMULA 01

Precision

`TP / (TP + FP)`

Proportion of positive identifications that were actually correct.

FORMULA 02

Recall

`TP / (TP + FN)`

Proportion of actual positives that were identified correctly.

FORMULA 03

F1 Score

`2 × P × R / (P + R)`

Harmonic mean balancing Precision and Recall.

Evaluation – Worked exact-match example

GROUND TRUTH

Reference Mentions

- **Anna Rossi** / PER **(Match)**
- **Acme Labs** / ORG
- **Milan** / LOC

SYSTEM OUTPUT

Predicted Mentions

- **Anna Rossi** / PER **(Exact Match)**
- **Labs** / ORG (Boundary Error)
- **Milan** / ORG (Type Error)

MATCHED

True Positives

TP = 1

Exact span & entity type match (Anna Rossi).

UNMATCHED PREDICTED

False Positives

FP = 2

Incorrect prediction boundaries or types.

MISSED REFERENCE

False Negatives

FN = 2

Unmatched ground truth entity mentions.

SCORES

P = R = F1

1/3 (33.3%)

1 / (1 + 2) = 0.33 exact match score

Evaluation – Why token accuracy can mislead

THE PROBLEM

Majority Class Dominance

- Many tokens may carry the O label.
- A system can achieve high token accuracy while finding no entities.

BEST PRACTICE

Entity-Level Assessment

- Report entity-level F1 and per-type results.
- Inspect boundary errors, type errors, and missed mentions.

CASE STUDY

All-O Prediction Baseline

Predicting every token as 'O' yields deceptively high accuracy due to class imbalance, despite failing completely at entity extraction.

TOKEN ACCURACY

90%

Deceptively High

ENTITY RECALL

0%

Zero Utility

Exercise – Annotation in pairs

TEXT TO ANNOTATE

"Anna Rossi from Acme Labs met Luca Bianchi in New York."

RULE 01

Entity Scheme

Use PER, ORG, and LOC with BIO2.

RULE 02

Tokenization

Treat punctuation as a separate token.

WORKFLOW

Pair Activity

Spend four minutes annotating, then compare with another pair.

Exercise – Solutions

Mention	First token	Second token
Anna Rossi	Anna / B-PER	Rossi / I-PER
Acme Labs	Acme / B-ORG	Labs / I-ORG
Luca Bianchi	Luca / B-PER	Bianchi / I-PER
New York	New / B-LOC	York / I-LOC

- The sentence contains four entity mentions.
- Each new mention begins with B-X.
- Each second word continues with I-X.

Modern approaches – NER through LLM prompting

- A prompt defines entity types and annotation rules.
- Zero-shot prompting supplies instructions without demonstrations.
- Few-shot prompting adds annotated examples.
- Generation can return spans in JSON or another structured format.

Modern Approaches

Structured Output & Verification

01 Extract only **PER, ORG**, and **LOC** mentions from input text.

02 Copy exact spans following classroom annotation guides.

03 Return a clean array of structured text and type pairs.

04 Check format validity & span correctness separately.

Output Schema Target

```
// Expected JSON Output Format
[
  {
    "text": "Anna Rossi",
    "type": "PER"
  },
  {
    "text": "Acme Labs",
    "type": "ORG"
  }
]
```

Related tasks

Entity linking and coreference

TASK 01

NER

NER assigns a type to a mention such as Milan.

TASK 02

Entity Linking

Entity linking maps that mention to a knowledge-base entry.

TASK 03

Coreference

Coreference groups mentions that refer to the same entity.

Each task requires additional decisions and evaluation.

Recap

Main ideas and questions

01

Span & Entity Typing

NER combines **span detection** with **entity typing**.

02

Sequence Labeling

BIO2 encodes flat mentions as a **token label sequence**.

03

Context & Guidelines

Context resolves ambiguity, while **guidelines** define the target.

04

Evaluation Metrics

Exact entity-level evaluation checks both **boundaries** and **types**.

Questions?

Comments?

Contacts

Dr. Georgios Peikos

georgios.peikos@unimib.it

Academic Year 2026/27

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Ikr3 Lab information and knowledge representation, retrieval, and reasoning LAB