Named Entity Recognition and Coreference Resolution in Natural Language Processing

Dr.Nourelhouda ZERARKA

Outline

1. Introduction
2. Entity Types and Applications
3. NER Approaches
4. Challenges & Solutions
5. Introduction to Coreference Resolution
6. Techniques of Coreference Resolution

What is Named Entity Recognition?

- Process of identifying and classifying named entities in text
- Named entities = real-world objects with proper names
- Example:
  "[Apple] is planning to open a new store in [Paris] next [March]"
  - Apple = ORGANIZATION
  - Paris = LOCATION
  - March = DATE

Why is NER Important?

- Information Extraction: detecting essential elements in the text, such as names of people, places, dates, and more.
- Search Engine Optimization: like Local and Contextual Search Optimization
- Question Answering Systems: identify the main focus of a user's question by recognizing entities
- Content Recommendation: NER allows recommendation systems to build detailed user profiles based on the entities they engage with.

Historical Context

- 1990s: Rule-based systems
- Early 2000s: Statistical methods (HMM, CRF)
- 2010s: Neural Networks
- Present: Transformer-based models

Common Entity Types

- PERSON: Names of people
- ORGANIZATION: Companies, institutions
- LOCATION: Cities, countries, addresses
- DATE/TIME: Temporal expressions
- MONEY: Monetary values
- PERCENT: Percentage values

Interactive Example

Identify entities in the following text:

"[John Smith] joined [Microsoft] in [2020] as CEO. The company's headquarters in [Redmond], [Washington] reported [$50 million] in profits last [quarter]."

- PERSON: John Smith
- ORG: Microsoft
- DATE: 2020, quarter
- LOC: Redmond, Washington
- MONEY: $50 million

Domain-Specific Entities

Biomedical

- GENE
- PROTEIN
- DISEASE
- DRUG

E-commerce

- PRODUCT
- BRAND
- CATEGORY
- PRICE

Computer science

- Software
- programming language
- hardware
- code

Rule-based Approach

- Uses pattern matching and dictionaries
- Pros:
  - Easy to implement
  - Interpretable results
  - No training data needed
- Cons:
  - Limited coverage
  - High maintenance
  - Rigid rules

Rule-based Example

```
import re

def find_dates(text):
    pattern = r'\d{1,2}/\d{1,2}/\d{4}'
    dates = re.findall(pattern, text)
    return dates

def find_emails(text):
    pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]
+\.[A-Z|a-z]{2,}\b'
    emails = re.findall(pattern, text)
    return emails
```

Machine Learning Approach

- Traditional ML algorithms (CRF, HMM, SVM)
- Features:
  - Word-level features
  - Context features
  - Part-of-speech tags
  - Gazetteer features
- Pros and Cons:
  - Better generalization than rules
  - Requires feature engineering
  - Needs annotated data

Deep Learning Approach

- Neural network architectures
  - BiLSTM-CRF
  - Transformer-based models
- Popular models:
  - BERT
  - RoBERTa
  - SpaCy
- Advantages:
  - No feature engineering
  - Better performance

Common Challenges

- Ambiguity: same name different entities
- Nested entities: named entities are contained within other named entities.
- Multi-word entities: Determining where multi-word entities start and end can be difficult.
- Domain adaptation: Different domains use specialized language and have unique entity types
- Out-of-vocabulary words: in social media, where slang, abbreviations, and new names emerge frequently

Ambiguity Examples

- "Washington"
  - PERSON: George Washington
  - LOCATION: Washington state
  - ORGANIZATION: Washington Post
- "Apple"
  - ORGANIZATION: Apple Inc.
  - PRODUCT: Apple iPhone
  - FOOD: apple fruit

What is Coreference Resolution?

- **Definition:** Finding all expressions that refer to the same entity in a text
- **Example:**

  Sarah opened the door. She was tired after work. The young doctor needed rest.

- All blue text refers to the same entity (Sarah)

Why is it Important?

- **Applications:**
  - Machine Translation: identify the gender or formality of entities in English and apply the correct pronouns in the target language.
  - Information Extraction: IE system can correctly identify and connect entities involved in relationships, even if they are mentioned in different sentences.
  - Question Answering: CR helps identify the correct antecedents, allowing the QA system to provide specific, accurate responses.
  - Text Summarization: summarization models can recognize when information about an entity is repeated under different names or pronouns. This allows the system to reduce redundancy, producing a more concise summary.

Types of Coreference

- **Pronominal**: when pronouns (he," "she," "they," "it") are used to refer back to a previously mentioned entity in the text.
  - "The cat chased the mouse. *It* was fast."
- **Nominal**: noun phrases that refer to the same entity but do not use pronouns. It uses general nouns or descriptions.
  - "Tim Cook leads Apple. *The man* announced new products."
- **Named Entity**: different mentions of the named entities refer to the same thing across a text.
  - "Microsoft released Windows 11. *The company* wants to sweep the world of IT."

Rule-Based Approaches

Key Rules:

- Number Agreement
  - ”The books... they...” (plural plural)
- Gender Agreement
  - ”Mary... she...” (feminine feminine)
- Person Agreement
  - ”I... me... my...” (1st person)
- Distance Heuristics (check for the closest noun phrase)
  - Sarah entered the room. She saw a **book** on the table. **It** looked interesting.”

Feature-Based Methods

Syntactic Features

- Grammatical role: Gender, Number, Person.
- Parse tree distance: Comparing syntactic positions.

Semantic Features

- Named entity type: “person,” “organization,” “location”
- Semantic roles “doctor” and “she” are both human

Machine Learning Approaches

Common Models:

- Mention-Pair Model
    - Classifies pairs as coreferent or not (each possible mention pair is evaluated for coreference, then the pairs with the highest probabilities are linked.)
- Entity-Mention Model
    - group multiple mentions together under the assumption that they refer to the same entity,
- Neural Models
    - End-to-end learning
    - Contextual embeddings

Implementation Example

```python
def resolve_coreference(text):
    # Step 1: Mention Detection
    mentions = detect_mentions(text)

    # Step 2: Feature Extraction
    features = extract_features(mentions)

    # Step 3: Classification
    pairs = create_mention_pairs(mentions)
    coreferent = classify_pairs(pairs, features)
```

Common Challenges

- World Knowledge Requirement
  - "The president met with his cabinet. The commander in chief..."
- Long-distance Dependencies
- Split Antecedents
  - "John met Mary. They went to dinner."
- Bridging Anaphora
  - "Sarah bought a new phone. The screen was already scratched"

Evaluation Metrics

- **MUC** (Message Understanding Conference)
  - Link-based metric
- **B-CUBED**
  - Mention-based metric
- **CEAF**
  - Entity-based metric
- **CoNLL F1**
  - Average of above metrics

Practical Example

Text Analysis:

John met Bill at his house. He offered him coffee.