CSCI E-89B Introduction to Natural Language Processing

Harvard Extension School

Dmitry Kurochkin

Fall 2025
Lecture 10

Contents

Contents

What is Named Entity Recognition?

- **Overview:** NER is a process for locating and classifying named entities in text into predefined categories such as:
  - Person names
  - Organizations (e.g., companies, institutions)
  - Geopolitical Entities (e.g., countries, cities)
  - Date and time expressions
  - Numerical expressions (e.g., percentages, monetary values)
  - Event names
  - Product names or brands
- **Applications:**
  - Information Extraction
  - Content Classification
  - Question Answering
  - Semantic Search
  - Data Annotation for Machine Learning

What is Named Entity Recognition? (Continued)

CHART: THE ECONOMIST

Markets continued to rally in response to Donald Trump's election victory. The S&P 500 hit another high (it has broken more than 50 records so far this year) and the Dow Jones Industrial Average closed above the 44,000 mark for the first time. The rise in Tesla's stock pushed the carmaker above a valuation of $1trn, which it last achieved in early 2022. The dollar continued to climb, hitting emerging-market currencies (over 40% of global trade is invoiced in dollars). Cryptocurrencies also made huge gains. Bitcoin surged by 30% in a week to trade above a record $90,000.

America's annual **inflation** rate rose for the first time since March. It stood at 2.6% in October, up from 2.4% in September. The core rate, which excludes volatile energy and food prices, held steady at 3.3%. Traders still expect the **Federal Reserve** to cut interest rates again next month. It recently shaved a quarter of a percentage point off its main rate, taking it to a range of between 4.5% and 4.75%. The Bank of England also reduced its rate by a quarter-point, to 4.75%.

Source: *The Economist.* (2024, November 14). Retrieved from `https://www-economist-com.ezp-prod1.hul.harvard.edu/the-world-this-week/2024/11/14/business`

What is Named Entity Recognition? (Continued)

Markets continued to rally in response to Donald Trump's PERSON election victory. The S&P 500 hit another high (it has broken more than 50 CARDINAL records so far this year DATE) and the Dow Jones Industrial Average closed above the 44,000 CARDINAL mark for the first ORDINAL time. The rise in Tesla ORG 's stock pushed the carmaker above a valuation of $ 1trn MONEY , which it last achieved in early 2022 DATE . The dollar continued to climb, hitting emerging-market currencies (over 40% PERCENT of global trade is invoiced in dollars). Cryptocurrencies ORG also made huge gains. Bitcoin PERSON surged by 30% PERCENT in a week DATE to trade above a record $ 90,000 MONEY . America GPE 's annual inflation rate rose for the first ORDINAL time since March DATE . It stood at 2.6% PERCENT in October DATE , up from 2.4% PERCENT in September DATE . The core rate, which excludes volatile energy and food prices, held steady at 3.3% PERCENT . Traders still expect the Federal Reserve ORG to cut interest rates again next month DATE . It recently shaved a quarter CARDINAL of a percentage point off its main rate, taking it to a range of between 4.5% and 4.75% PERCENT . The Bank of England ORG also reduced its rate by a quarter CARDINAL -point, to 4.75% PERCENT .

Contents

Applications of NER in NLP

- **Enhancing Text Classification:**
  - Feature Enhancement: Utilize named entities as features (e.g., companies, locations), enriching classifier inputs.
  - Dimensionality Reduction: Focus on core entities to minimize noise and reduce dimensional space, improving model efficiency.
  - Contextual Understanding: Recognize relationships (e.g., "Apple" and "Tim Cook") to refine context-aware classifications.
- **Information Extraction:**
  - Structured Data Creation: Extract entities to populate structured databases, aiding in metadata development.
- **Content and Semantic Annotation:**
  - Tagging for Categorization: Use NER to tag content, enhancing categorization and retrieval.
  - Enhanced Text Annotation: Auto-tag documents with entities, enabling hierarchical classification paths.

Applications of NER in NLP (Continued)

- **Question Answering:**
  - Linking for Precision: Identify and link entities in questions and source texts to improve the relevance and accuracy of retrieved answers.
- **Contextual and Sentiment Analysis:**
  - Semantic Analysis: Detect key entities for thematic analysis and associate sentiment to specific entities, offering insightful opinions.
- **Hybrid Approaches:**
  - Combining Techniques: Integrate NER outcomes with rule-based filters and machine learning algorithms to enhance overall accuracy and performance.

Contents

Rule-based Methods for NER: Overview and Techniques

- **Overview:**
  - Rule-based Named Entity Recognition (NER) uses a collection of hand-crafted linguistic rules to detect and classify named entities in text.
  - Techniques include:
    - ★ **Regular Expressions:** Patterns for matching text strings.
    - ★ **Grammar Rules:** Linguistic rules that define how words can be combined to form valid structures that help in identifying entities.
    - ★ **Gazetteers:** Predefined lists of known entities, such as city names, companies, and organizations, used for direct lookup.
    - ★ **Syntactic Patterns:** Structural patterns in sentences that help detect entity boundaries and relationships.
    - ★ **Heuristic Methods:** Simple, rule-based approaches that apply logical reasoning based on context and common usage.
    - ★ **Semantic Analysis:** Techniques that involve understanding the meaning of the words in context to improve entity recognition.
    - ★ **Contextual Indicators:** Use of surrounding word patterns or phrases that often precede or follow a named entity.
  - Commonly implemented via Finite State Machines or dedicated rule engines.

Rule-based Methods for NER: Pros and Cons

- **Pros:**
  - **High Precision:** Particularly effective in well-defined domains where rules can capture entities accurately.
  - **Human Interpretability:** Rules are explicit and understandable, facilitating easier interpretation and modification.
  - **No Need for Large Labeled Datasets:** Operates efficiently without extensive training data, unlike machine learning approaches.
- **Cons:**
  - **Limited Flexibility:** Challenging to adapt to new languages or domains without manual adjustment.
  - **Domain Expertise Requirement:** Requires extensive domain knowledge to create comprehensive rules, which can be time-consuming and costly.
  - **Maintenance Burden:** Requires regular updates as language use evolves and new text types emerge.
  - **Scalability Issues:** Rule sets can become complex and hard to manage as they grow.

Contents

Rule-based NER in Python: Example

```
import re

def detect_entities(text):
    entities = {
        'Dates': [],
        'Emails': [],
        'Times': [],
        'Titles': [],
        'Organizations': [],
    }

    # Regular expressions
    date_pattern = r'\b\d{1,2}/\d{1,2}/\d{4}\b|' \
                   r'\b(?:January|February|March|April|May|June|July|August|September|October|\
                   November|December) ' \
                   r'(?:\d{1,2},\s*)?\d{4}\b'
    email_pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{2,}\b'
    time_pattern = r'\b\d{1,2}:\d{2}\s*(?:AM|PM)\b'
    title_pattern = r'\bDr\.\s+[A-Z][a-z]+\b'
    org_pattern = r'\b(?:[A-Z][a-zA-Z]+(?:\s+[A-Z][a-zA-Z]+)*)\s+(?:Inc|Ltd|Corp|\
    Incorporated)\b'

    # Finding matches
    entities['Dates'] = re.findall(date_pattern, text)
    entities['Emails'] = re.findall(email_pattern, text)
    entities['Times'] = re.findall(time_pattern, text, flags=re.IGNORECASE)
    entities['Titles'] = re.findall(title_pattern, text)
    entities['Organizations'] = re.findall(org_pattern, text)

    return entities
```

Rule-based NER in Python: Example (Continued)

```
text = """
Dr. John Smith met with representatives from Apple Inc. and Google LLC at their
headquarters in New York City to discuss their new product launch on November 18, 2024.
His email is john.smith@email.com, and the meeting was scheduled to start at 10:00 AM
sharp. Later in the afternoon, Dr. Emily White was expected to join them, providing
valuable insights on the project. Coincidentally, that day marked the 5th anniversary
of their partnership, which began in November 2019. After the meeting, the team dined
at a famous rooftop restaurant overlooking the Statue of Liberty.
"""


# Execute function and display results
entities_found = detect_entities(text)
for entity_type, values in entities_found.items():
    print(f"{entity_type}: {values}")
```

```
Dates: ['November 18, 2024', 'November 2019']
Emails: ['john.smith@email.com']
Times: ['10:00 AM']
Titles: ['Dr. John', 'Dr. Emily']
Organizations: ['Apple Inc']
```

Contents

Statistical Methods for NER: Overview

- **Overview:**
  - Statistical methods apply probabilistic models to determine the likelihood of sequences of labels, enabling the identification and classification of entities within text data.
- **Common Techniques:**
  - **Hidden Markov Models (HMMs):** Use observable data to predict sequences of hidden states, particularly useful in simpler NER applications.
  - **Conditional Random Fields (CRFs):** A discriminative, sequential modeling technique that excels at label prediction by modeling dependencies between observation sequences.
  - **Neural Networks (e.g., CNN, LSTMs, BERT):** Leveraging deep learning to capture complex patterns and dependencies, these methods achieve high accuracy, especially with contextualized embeddings such as in BERT.

Contents

Statistical Methods for NER: Advantages and Challenges

- **Advantages:**
  - **Data-Driven Learning:** Capable of learning patterns directly from data, reducing the need for manually crafted rules.
  - **Domain Flexibility:** Adaptable to different domains with appropriate training data, improving performance in versatile contexts.
- **Challenges:**
  - **Data Requirements:** Require sufficiently large and well-annotated datasets to train effectively, which can be resource-intensive to collect and maintain.
  - **Computational Demand:** High computational resources are often needed, especially for deep learning models, which can limit accessibility and scalability.

Contents

Using SpaCy for NER

```
# Installing SpaCy and its language model
# !pip install spacy
# !python -m spacy download en_core_web_sm

import spacy
from spacy import displacy

# Load the SpaCy model
nlp = spacy.load("en_core_web_sm")

# Example text
text = "Markets continued to rally in response to Donald Trump's election victory. The
S&P 500 hit another high (it has broken more than 50 records so far this year) and the
Dow Jones Industrial Average closed above the 44,000 mark for the first time. The rise
in Tesla's stock pushed the carmaker above a valuation of $1trn, which it last achieved
in early 2022. The dollar continued to climb, hitting emerging-market currencies (over
40% of global trade is invoiced in dollars). Cryptocurrencies also made huge gains.
Bitcoin surged by 30% in a week to trade above a record $90,000. America's annual
inflation rate rose for the first time since March. It stood at 2.6% in October, up from
2.4% in September. The core rate, which excludes volatile energy and food prices, held
steady at 3.3%. Traders still expect the Federal Reserve to cut interest rates again
next month. It recently shaved a quarter of a percentage point off its main rate, taking
it to a range of between 4.5% and 4.75%. The Bank of England also reduced its rate by a
quarter-point, to 4.75%."

# Process the text with SpaCy's NLP pipeline
doc = nlp(text)
```

Using SpaCy for NER (Continued)

```
# Visualize the entities in the text using displacy
displacy.render(doc, style="ent", jupyter=True)
```

Markets continued to rally in response to Donald Trump's PERSON election victory. The S&P 500 hit another high (it has broken more than 50 CARDINAL records so far this year DATE) and the Dow Jones Industrial Average closed above the 44,000 CARDINAL mark for the first ORDINAL time. The rise in Tesla ORG 's stock pushed the carmaker above a valuation of $ 1trn MONEY , which it last achieved in early 2022 DATE . The dollar continued to climb, hitting emerging-market currencies (over 40% PERCENT of global trade is invoiced in dollars). Cryptocurrencies ORG also made huge gains. Bitcoin PERSON surged by 30% PERCENT in a week DATE to trade above a record $ 90,000 MONEY . America GPE 's annual inflation rate rose for the first ORDINAL time since March DATE . It stood at 2.6% PERCENT in October DATE , up from 2.4% PERCENT in September DATE . The core rate, which excludes volatile energy and food prices, held steady at 3.3% PERCENT . Traders still expect the Federal Reserve ORG to cut interest rates again next month DATE . It recently shaved a quarter CARDINAL of a percentage point off its main rate, taking it to a range of between 4.5% and 4.75% PERCENT . The Bank of England ORG also reduced its rate by a quarter CARDINAL -point, to 4.75% PERCENT .

Using SpaCy for NER (Continued)

```
# Access and print the details of each entity
for ent in doc.ents:
    print(f"Entity Text: {ent.text} | Label: {ent.label_} |
          Index Range: {ent.start_char}-{ent.end_char}")
```

```
Entity Text: Donald Trump's | Label: PERSON | Index Range: 42-56
Entity Text: more than 50 | Label: CARDINAL | Index Range: 119-131
Entity Text: this year | Label: DATE | Index Range: 147-156
Entity Text: 44,000 | Label: CARDINAL | Index Range: 212-218
Entity Text: first | Label: ORDINAL | Index Range: 232-237
Entity Text: Tesla | Label: ORG | Index Range: 256-261
Entity Text: 1trn | Label: MONEY | Index Range: 312-316
Entity Text: early 2022 | Label: DATE | Index Range: 344-354
Entity Text: over 40% | Label: PERCENT | Index Range: 423-431
Entity Text: Cryptocurrencies | Label: ORG | Index Range: 473-489
Entity Text: Bitcoin | Label: PERSON | Index Range: 512-519
Entity Text: 30% | Label: PERCENT | Index Range: 530-533
Entity Text: a week | Label: DATE | Index Range: 537-543
Entity Text: 90,000 | Label: MONEY | Index Range: 569-575
Entity Text: America | Label: GPE | Index Range: 577-584
Entity Text: first | Label: ORDINAL | Index Range: 622-627
Entity Text: March | Label: DATE | Index Range: 639-644
Entity Text: 2.6% | Label: PERCENT | Index Range: 658-662
Entity Text: October | Label: DATE | Index Range: 666-673
Entity Text: 2.4% | Label: PERCENT | Index Range: 683-687
Entity Text: September | Label: DATE | Index Range: 691-700
Entity Text: 3.3% | Label: PERCENT | Index Range: 780-784
Entity Text: the Federal Reserve | Label: ORG | Index Range: 807-826
Entity Text: next month | Label: DATE | Index Range: 855-865
Entity Text: a quarter | Label: CARDINAL | Index Range: 886-895
Entity Text: between 4.5% and 4.75% | Label: PERCENT | Index Range: 961-983
Entity Text: The Bank of England | Label: ORG | Index Range: 985-1004
Entity Text: quarter | Label: CARDINAL | Index Range: 1032-1039
Entity Text: 4.75% | Label: PERCENT | Index Range: 1050-1055
```