Entities

Named Entity Recognition (NER)

- This is the task of identifying and classifying named entities in a text.
- Named entities are names of persons, organizations, locations, dates, etc.
- This is typically approached as a sequence labeling task, where each word in a sentence is labeled with the type of named entity it represents.

Named Entity Recognition (NER)

- Task was first introduced at the Message Understanding Conference (MUC) in 1995.
- Most versions identify proper nouns, such as names of people, places, and organizations.
- Some also identify temporal expressions, such as dates and times.
- Domain-specific versions exist such as GENIA for biomedical text, which has 36 classes!

CoNLL 2003 Dataset

- This dataset is a popular benchmark for NER.
- It contains news articles from Reuters, and is annotated with named entities.
- The entities are divided into four categories: persons, organizations, locations, and miscellaneous.

IOB Annotation

- In the CoNLL 2003 dataset, named entities are annotated using the IOB format.
    - I: Inside a named entity
    - O: Outside a named entity
    - B: Beginning of a named entity

Example

- Consider the sentence: “Lady Gaga is from New York.”

- The IOB annotation for this sentence is:
  - Lady B-PER
  - Gaga I-PER
  - is O
  - from O
  - New B-LOC
  - York I-LOC

Methods for Named Entity Recognition

- Rule-based methods:
  - Rule-based approach rely on hand-crafted rules to identify named entities.
  - Use domain specific gazetteers (list of names)
  - Combined with regular expressions
- Unsupervised methods:
  - Use syntactic patterns to identify entities
  - Clustering of words based on context

Feature-based Approaches

- Use features such as:
  - Part-of-speech tags
  - Word embeddings
  - Contextual embeddings
  - Word shapes
  - Capitalization
  - Prefixes and suffixes
  - etc.
- Solved with "Good Old Machine Learning" (e.g., SVM)

Deep Learning Approach

- Input features are extracted for each word:
  - Word-level features (e.g., word embeddings)
  - Character-level features (often using CNNs)
  - Hybrid features (e.g., gazeteers, lexical similarity or POS tags)
- As named entities often use unusual words, character-level features are helpful

---

Encoding and Decoding

- Context Encoder:
  - A model (e.g., BiLSTM, Transformer) is used to encode the context of each word.
- Tag Decoder:
  - Probabilities of tags can be calculated using a softmax layer.
  - The most likely sequence of tags can be predicted with a Viterbi decoder.
  - Conditional Random Fields (CRF) can also be used to model the dependencies between tags.

Evaluation

- Precision, Recall, and F1-score are used to evaluate NER systems.
- This is done with “exact-match evaluation”
  - A named entity is considered correct only if its boundaries are correctly predicted.
- Some evaluations (e.g. MUC-6) used partial match, however this makes evaluation unintuitive and is not commonly used.

Automated Term Extraction

- Automated term extraction is the task of identifying terms (e.g., multi-word expressions) in a text.
- Terms are domain-specific, and can be used to build a domain-specific ontology.
- Named entities are often terms, but not all terms are named entities.

Key-phrase Extraction

- Key-phrase extraction is the task of identifying important phrases in a text.
- These are typically used to summarize and index a document.
- Key-phrases are not necessarily terms or named entities, and can be single words or multi-word expressions.

Example

> Within the cutting-edge oncology department of St. Mary’s Hospital, groundbreaking research is underway, focusing on innovative

treatments for aggressive brain tumors like glioma.

- Named entities: St. Mary's Hospital
- Terms: oncology, brain tumors, glioma
- Key-phrases: groundbreaking research, innovative treatments

---

Automated Term Extraction

- Unsupervised methods for term extraction:
  - Use statistical measures such as frequency, mutual information, or pointwise mutual information.
  - Use syntactic patterns to identify terms.
- The state-of-the-art approaches for term extraction are similar to NER:
  - Deep-learning methods using pretrained language models (e.g., RoBERTa)

Key-phrase Extraction

- Unsupervised methods are more successful for key-phrase extraction.

KeyBERT

- N-grams are extracted from the text.
- For each n-gram, BERT is used to calculate the embedding.
- An embedding is also calculated for the entire document.
- The cosine similarity between the n-gram embedding and the document embedding is used to rank the n-grams.
- The top n-grams are selected as key-phrases.

Entity Linking

- Entity linking is the task of linking named entities to a knowledge base.
- Typically a knowledge base such as Wikipedia or Wikidata is used.
- This is useful for disambiguating named entities, and for enriching the text with additional information.

Entity Linking - Example

> **Swift** began writing songs professionally at age 14 and signed with
> Big Machine Records in 2005.

Candidate entities for Swift: * Taylor Swift (singer) * Swift (programming language) * Society for Worldwide Interbank Financial Telecommunication (SWIFT)

Entity Linking

- Entity linking is typically approached as a multi-step process:
  - Named entity recognition
  - Candidate generation
  - Entity Ranking
- State-of-the-art approaches create an embedding for the word in context and an embedding for each candidate.

Candidate Generation

- Candidate generation is the task of generating a list of candidate entities for a named entity.
- Entities are found in a knowledge base, such as Wikipedia or Wikidata.
- Mentions of entities may be shortened.
- Entities may also have aliases (e.g., "The Beatles" and "The Fab Four").

Context embedding

- The context embedding is generally the word embedding of the named entity in the context.
  - Methods like attention or pooling are used to combine the word embeddings of multiword entities.

Candidate embedding

- The candidate embedding is dependent on the knowledge base:
  - First sentence of the Wikipedia page
  - Graph neural network for a knowledge graph
- Similarity is calculated using a Siamese network or cosine similarity.

Summary

- Named Entity Recognition (NER) is the task of identifying and classifying named entities in a text.

- Named entities are typically names of persons, organizations, locations, dates, etc.
- NER is typically approached as a sequence labeling task.
- Automated term extraction and key-phrase extraction are related tasks, but are not the same as NER.
- Entity linking is the task of linking named entities to a knowledge base.