Named Entity Recognition and Relation Detection for Biomedical Information Extraction

Authors: Nadeesha Perera, Matthias Dehmer and Frank Emmert-Streib

Mariana Chaves

UNIVERSITÉ
CÔTE D'AZUR

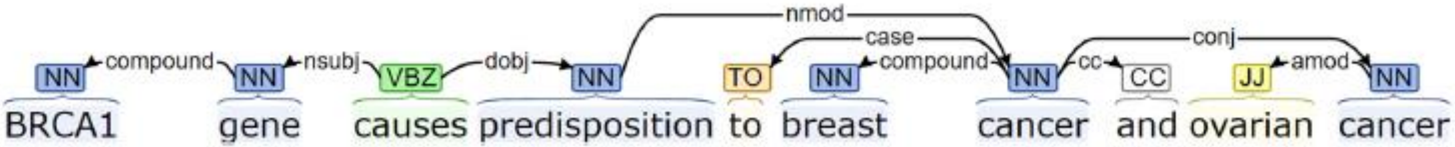
Introduction

BioNER : Biomedical Named Entity Recognition

Named Entity Recognition involves the automatic scanning through unstructured text to locate **“entities,”** for term normalization and classification into categories (genes, proteins, diseases…)

BioRD: Biomedical Relation Detection

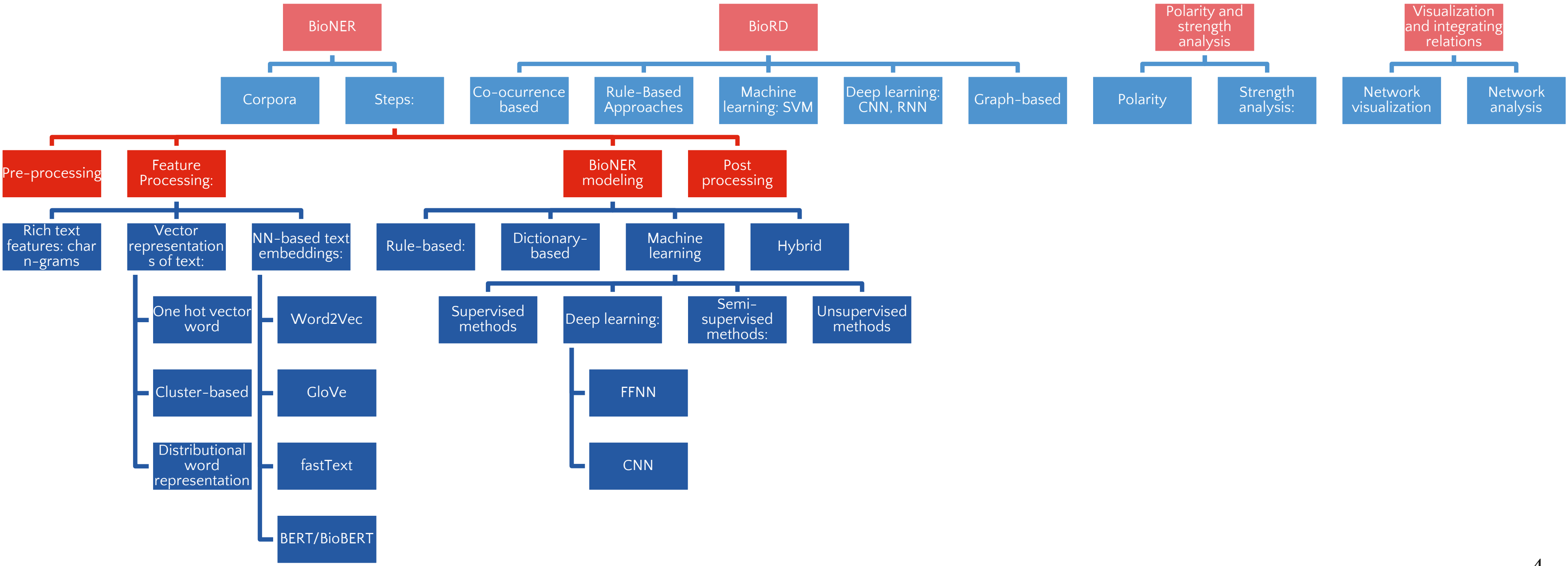
Connect biomedical entities to find meaningful **interactions.**



Challenges

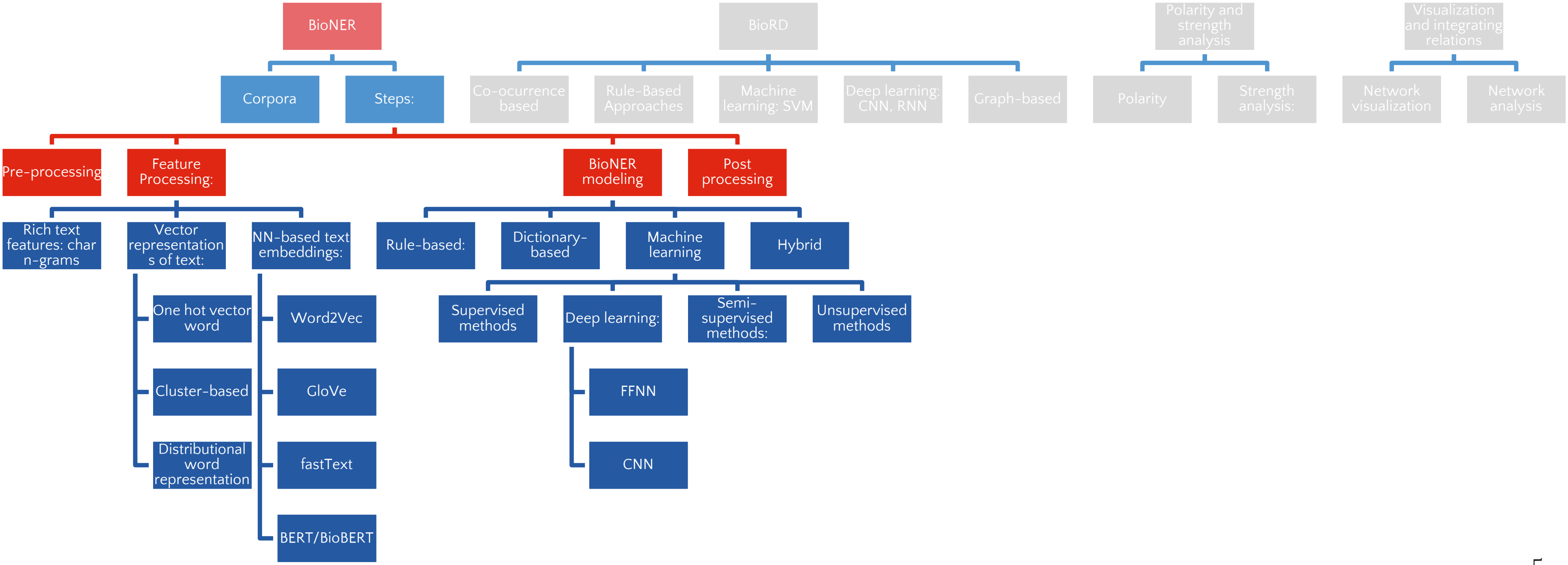
- Increasing **number of papers** in the field.
- General NER models were not made for medical terms
- Non-standard use of **abbreviations** (CLD, could either refer to "Cholesterol-lowering Drug," "Chronic Liver Disease,"), synonyms, homonyms, polysemy, non-standard names ("Lymphocytic Leukemia", and "Lymphoblastic Leukemia"), long chain words (epidemic transient diaphragmatic spasm)

The general landscape



- BioNER
  - Corpora
  - Steps:
    - Pre-processing
    - Feature Processing:
      - Rich text features: char n-grams
      - Vector representations of text:
        - One hot vector word
        - Cluster-based
        - Distributional word representation
      - NN-based text embeddings:
        - Word2Vec
        - GloVe
        - fastText
        - BERT/BioBERT
    - BioNER modeling
      - Rule-based:
      - Dictionary-based
      - Machine learning
        - Supervised methods
        - Deep learning:
          - FFNN
          - CNN
        - Semi-supervised methods:
        - Unsupervised methods
      - Hybrid
    - Post processing
- BioRD
  - Co-ocurrence based
  - Rule-Based Approaches
  - Machine learning: SVM
  - Deep learning: CNN, RNN
  - Graph-based
- Polarity and strength analysis
  - Polarity
  - Strength analysis:
- Visualization and integrating relations
  - Network visualization
  - Network analysis

The general landscape



BioNER

Corpora
GENETAG, JNLPBA, BioCreative corpora, GENIA, CRAFT

Pre-processing
Data cleaning, tokenization, stopwords, stemming, lemmatization, spelling correction…

Feature processing
Transform text into real-value word representations

BioNer modeling
Recognizing the entities

Post-processing
Resolving abbreviation ambiguities, disambiguation of classes and terms, coreferences (anaphoras)

BioNER – Feature processing

transform text into real-value word representations

01 Rich text features

char n-grams

02 Vector representations of text

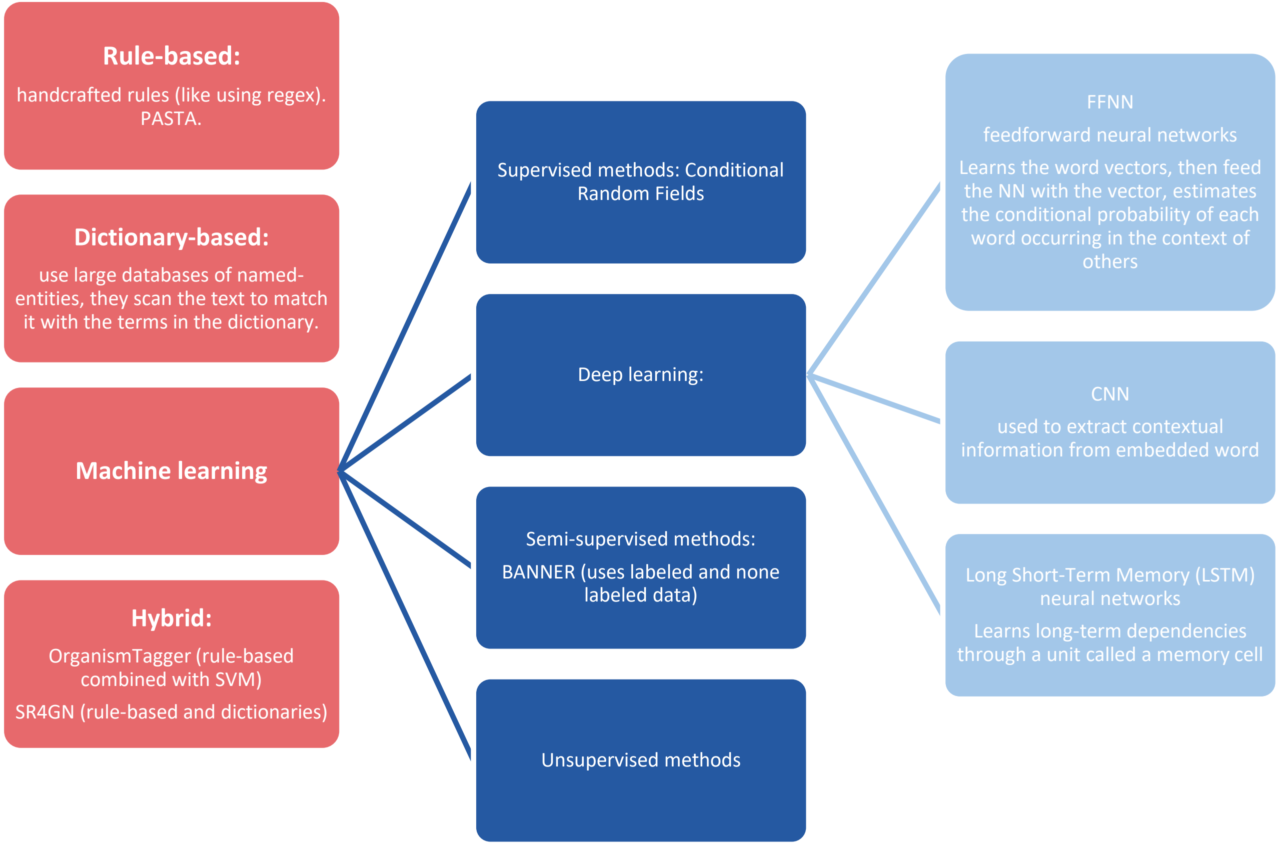
- **One hot vector word**
- **Cluster-based:** each cluster of words contains words with contextually similar information. Most famous is Brown clustering: hierarchical, similar paths and similar parents among words indicates close semantics/relationships
- **Distributional word representation:** uses co-ocurrence matrices

03 Neural network-based text embeddings

- **Word2Vec:** 2 layer NN, takes a corpus, creates a vocabulary, produces uni-dimensional vectors, creates a vector space were similar words are close to each other. Two possible algorithms: Continous Bag-of-Words, Continuous Skip-Gram.
- **GloVe:** global corpus-wide statistics are captured by the method
- **fastText:** Instead of directly learning the vector representation of a word, it first learns the word as a representation of N-gram characters. Very effective in representing suffixes/prefixes, and the embedding of rare words)
- **BERT/BioBERT**: uses the transformer learning model to learn contextual token embeddings of a given sentence bidirectionally

BioNER – BioNer modeling

Recognizing the entities



Rule-based:

handcrafted rules (like using regex). PASTA.

Dictionary-based:

use large databases of named-entities, they scan the text to match it with the terms in the dictionary.

Machine learning

Hybrid:

OrganismTagger (rule-based combined with SVM)

SR4GN (rule-based and dictionaries)

Supervised methods: Conditional Random Fields

Deep learning:

Semi-supervised methods:

BANNER (uses labeled and none labeled data)

Unsupervised methods

FFNN

feedforward neural networks

Learns the word vectors, then feed the NN with the vector, estimates the conditional probability of each word occurring in the context of others

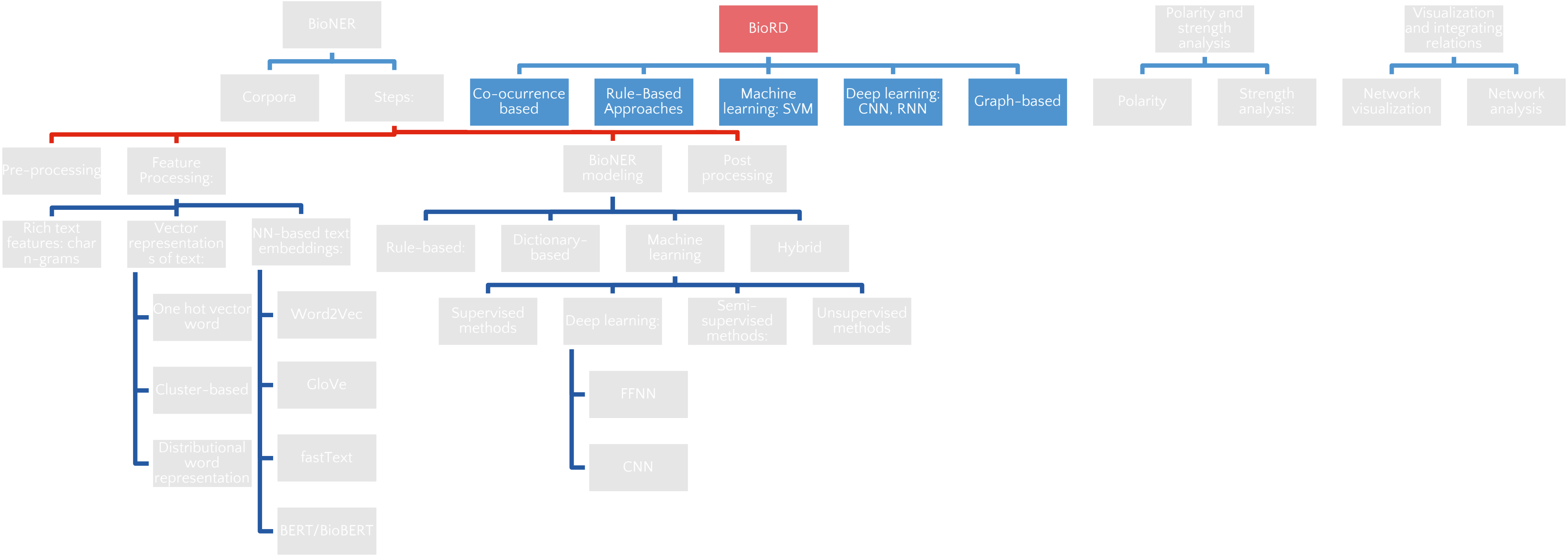
CNN

used to extract contextual information from embedded word

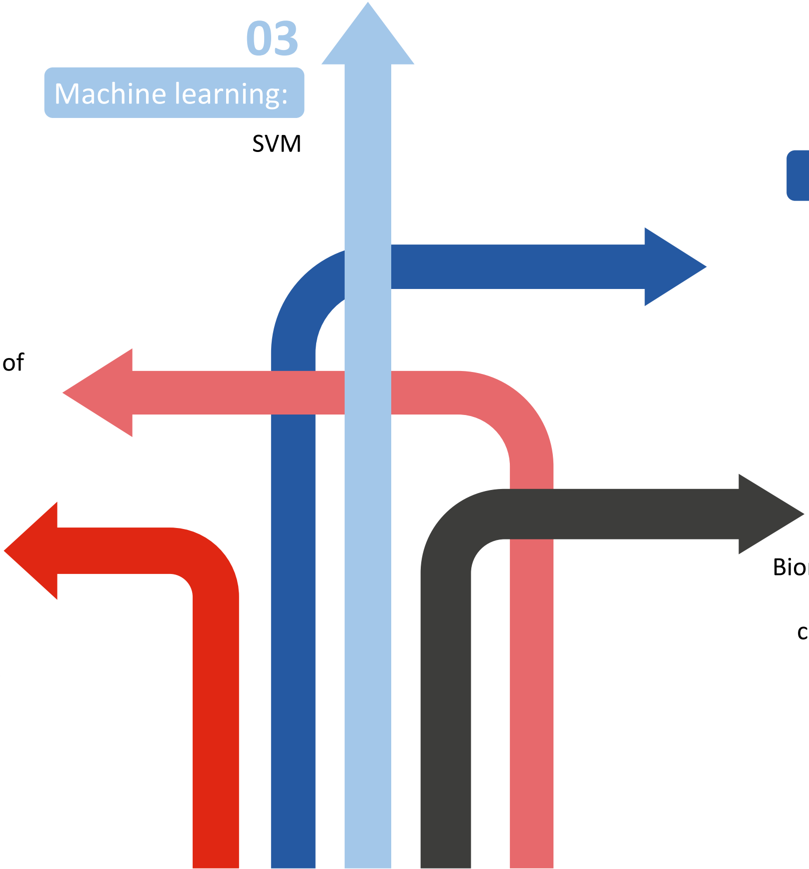
Long Short-Term Memory (LSTM) neural networks

Learns long-term dependencies through a unit called a memory cell

The general landscape



BioRD



01

Co-ocurrence based:

The hypothesis is that the more frequent two entities occur together, the higher the probability that they are associated with each other.

02

Rule-Based Approaches

They Rely on part-of-speech (POS) tagging tools to identify associations, by scanning for verbs and prepositions that correlate two or more nouns. List of verbs that are considered to show implications between nouns: catalyzes, influences, mutates.

.

03

Machine learning:

SVM

04

Deep learning

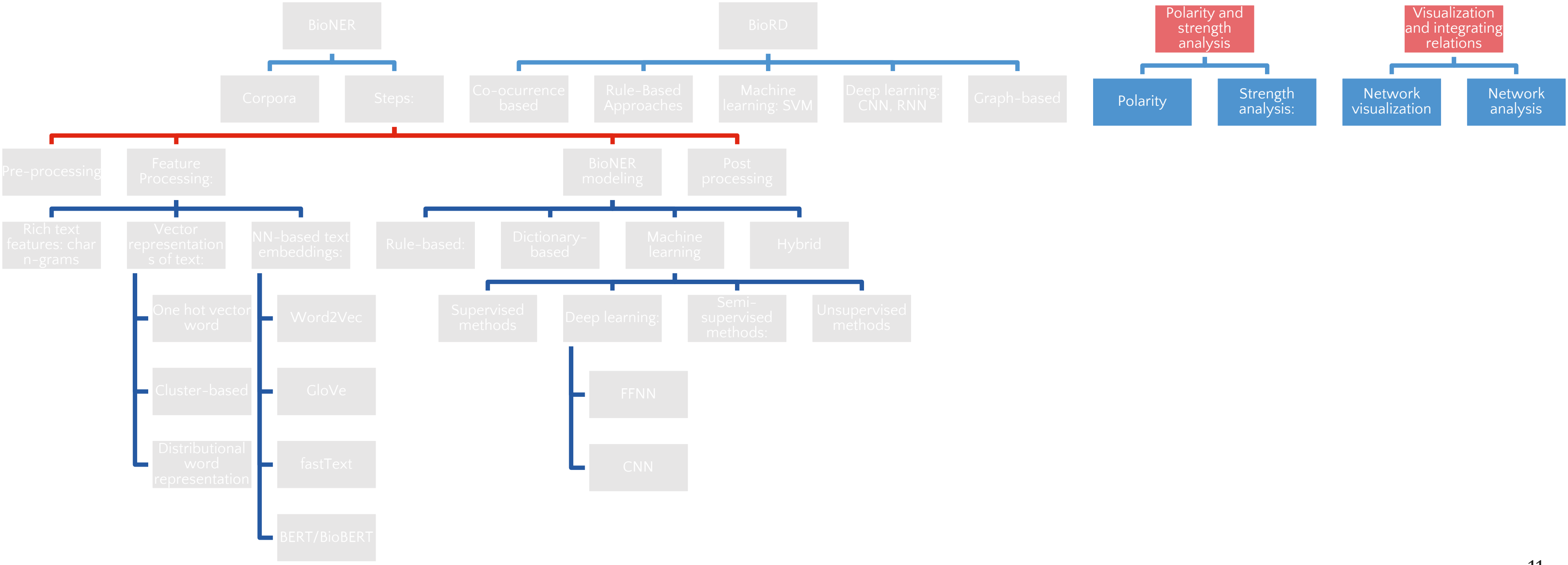
CNN RNN

05

Graph-based

Biomedical named entities are vertices and other syntactic/semantic structures connecting them are edges. This method facilitates identifying common syntactic patterns.

The general landscape



- BioNER
  - Corpora
  - Steps:
    - Pre-processing
    - Feature Processing:
      - Rich text features: char n-grams
      - Vector representations of text:
        - One hot vector word
        - Cluster-based
        - Distributional word representation
      - NN-based text embeddings:
        - Word2Vec
        - GloVe
        - fastText
        - BERT/BioBERT
    - BioNER modeling
      - Rule-based:
      - Dictionary-based
      - Machine learning
        - Supervised methods
        - Deep learning:
          - FFNN
          - CNN
        - Semi-supervised methods:
        - Unsupervised methods
      - Hybrid
    - Post processing
- BioRD
  - Co-ocurrence based
  - Rule-Based Approaches
  - Machine learning: SVM
  - Deep learning: CNN, RNN
  - Graph-based
- Polarity and strength analysis
  - Polarity
  - Strength analysis:
- Visualization and integrating relations
  - Network visualization
  - Network analysis

Polarity and Strength – Visualization

Polarity and strength analysis

Polarity

Similar to sentiment analysis it identifies positive, negative or neutral associations.

Strength analysis

After identifying associations between entities we want to measure how strong is the relationship. Polysearch, syntactic parse trees

Visualization and integrating relations

Network visualization:

Nodes (also called vertices) correspond to entities and edges (also called links) to relations between entities. Cytoscape, Gephi, NetbioV (R), Graph-tool (Python)

Network analysis:

Node centrality measures, shortest paths (centrality measures are commonly used to identifying the importance of an entity within the entire network), network clustering, and network density (compares the number of existing relations between the nodes vs. all possible connections that can be formed in the network)