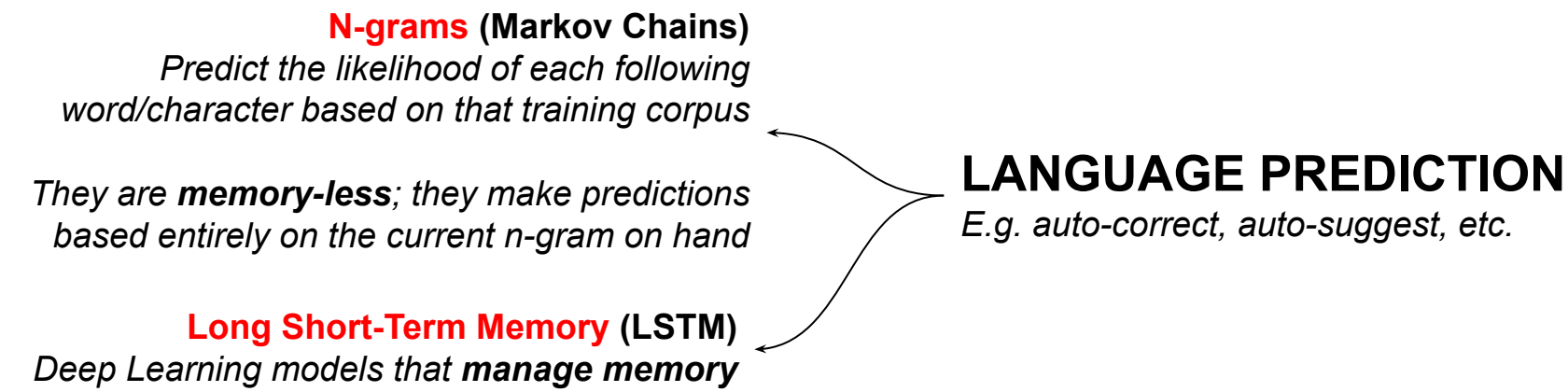
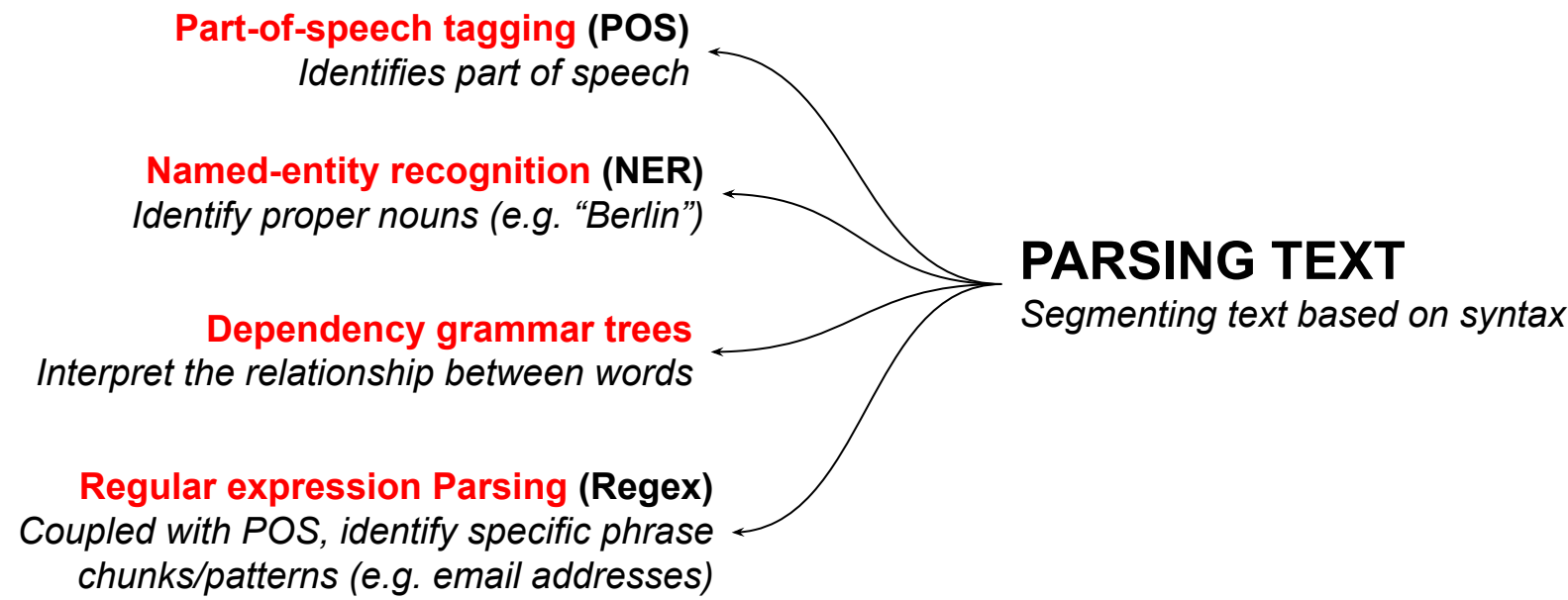
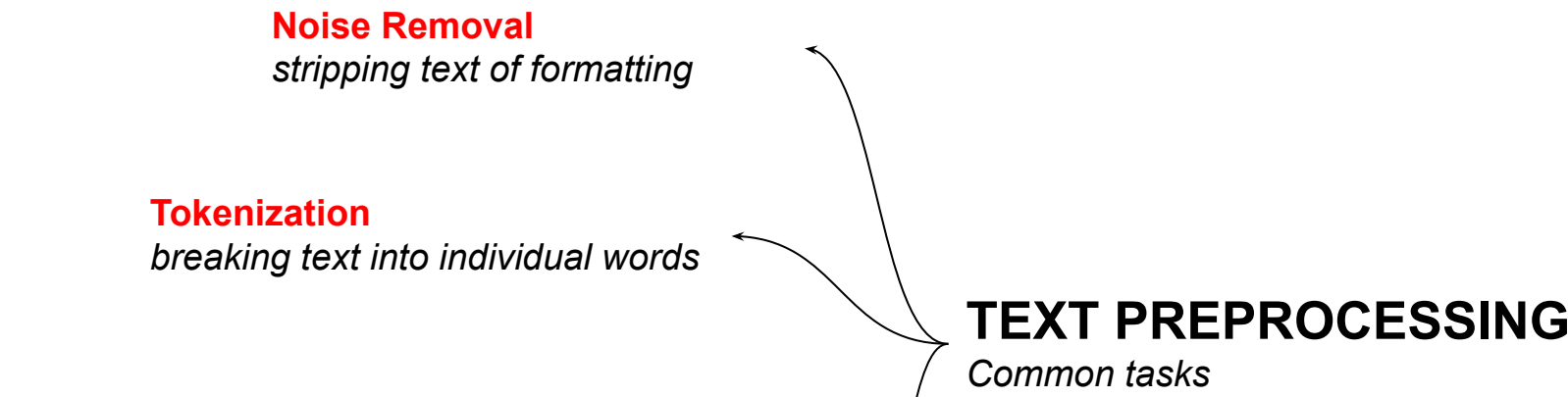


# NATURAL LANGUAGE PROCESSING

intersection of linguistics, AI, and computer science



## CONSIDERATIONS:

- Different languages (**cultural & linguistic biases**)
- NLP can limit, but also **propagate bias** (within code or corpus)
- Privacy** issues

## LANGUAGE MODELS

Probabilistic computer models of language

- Bag-of-Words (BoW)**
  - Unigram model which tally count each word instance
  - Suitable for making predictions concerning **text topic or sentiment**
  - When **grammar & word order are irrelevant**
- N-gram**  
Considers a sequence of *n* units & calculates the probability of each unit in a body of language given the preceding sequence of length *n*
- Neural Language Models (NLMs)**  
**Deep Learning approach**, e.g. LSTMs, transformer models, etc.

## TOPIC MODELLING

An area of NLP dedicated to uncovering latent (hidden) topics within a body of language

- Term Frequency-Inverse Document Frequency (TF-IDF)**  
Deprioritize the most common words and **prioritize less frequently terms**
- Latent Dirichlet Allocation (LDA)**  
The next step after Bow or TF-IDF; a statistical model that takes the docs and determines **which word keeps popping up together in the same contexts** (documents)
- Word2Vec**  
Maps out the topic model results spatially as vectors so that **similarly used words are closer together (word embedding)**

## TEXT SIMILARITY

- Levenshtein distance** (minimal edit distance)  
The minimum number of insertions, deletions, and substitution that would need to occur for **one word to become another**
- Phonetic similarity**  
How much two words/phrases **sound the same**
- Lexical similarity**  
The degree to which texts use the **same vocabulary & phrases**
- Semantic similarity**  
The degree to which documents contain **similar meaning or topics** (recommendation systems)

## ADVANCED NLP TOPICS:

- Machine **translation** (NNs, LSTMs)
- Bias** Detection
- Language **Accessibility** (text2speech, speech recognition)