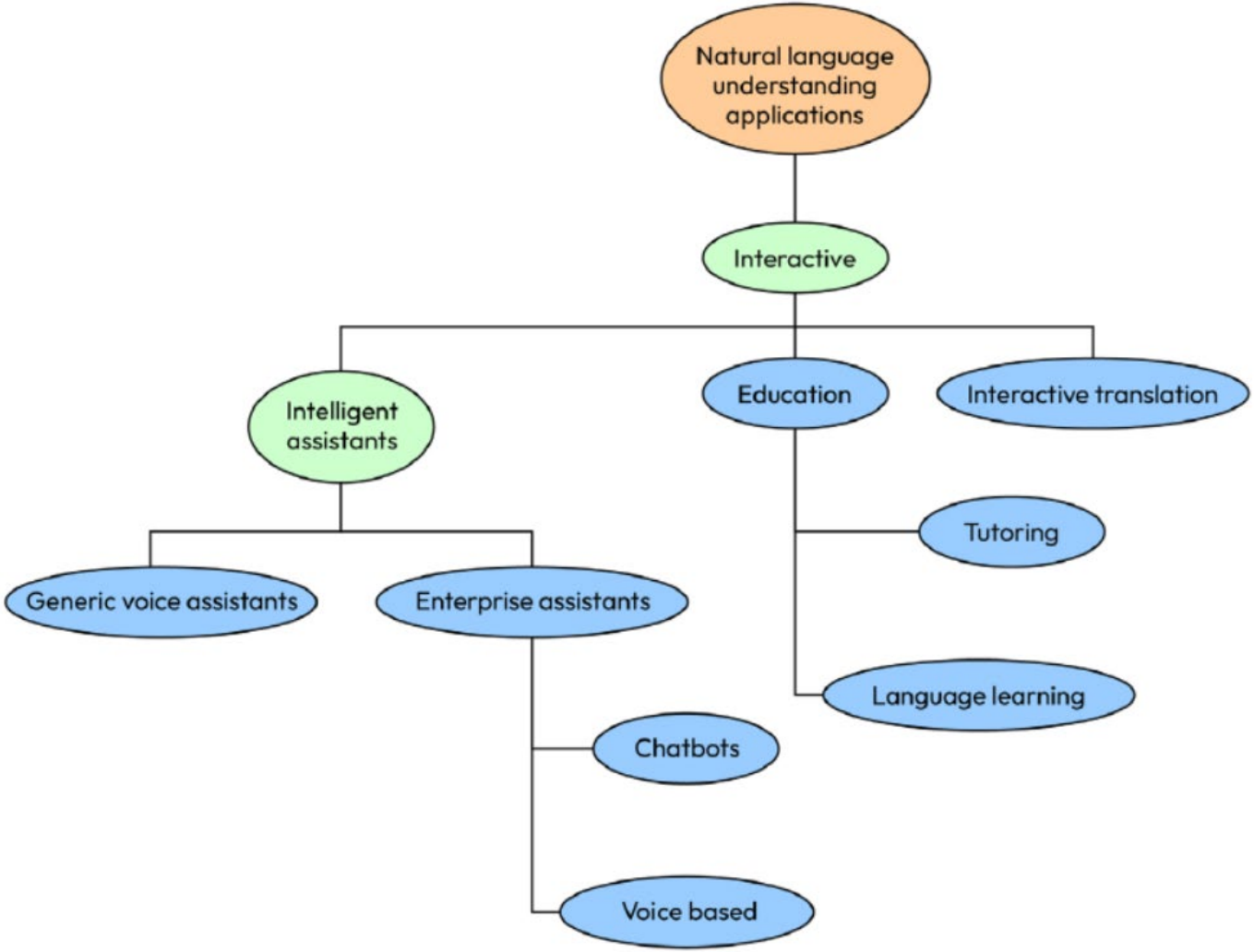
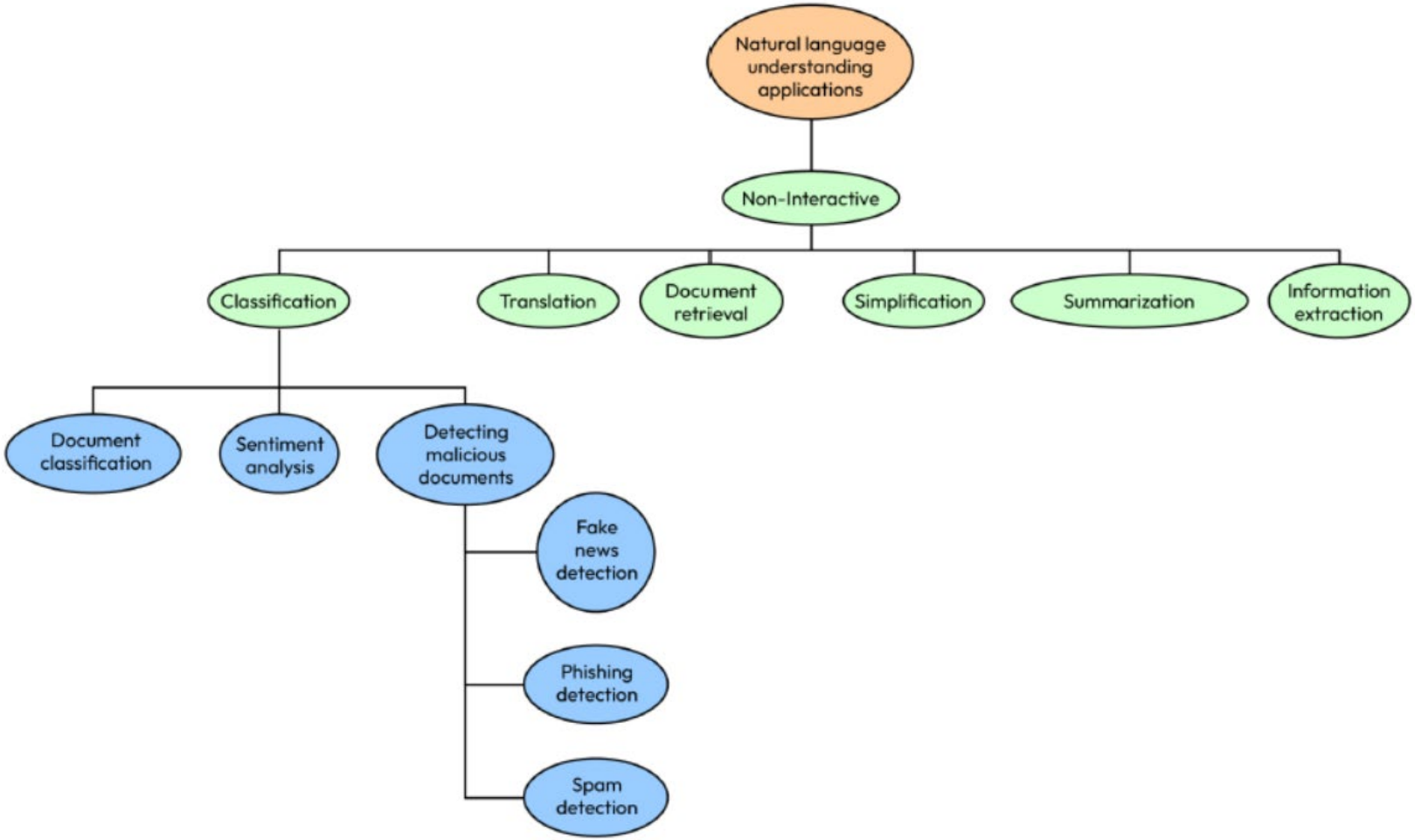
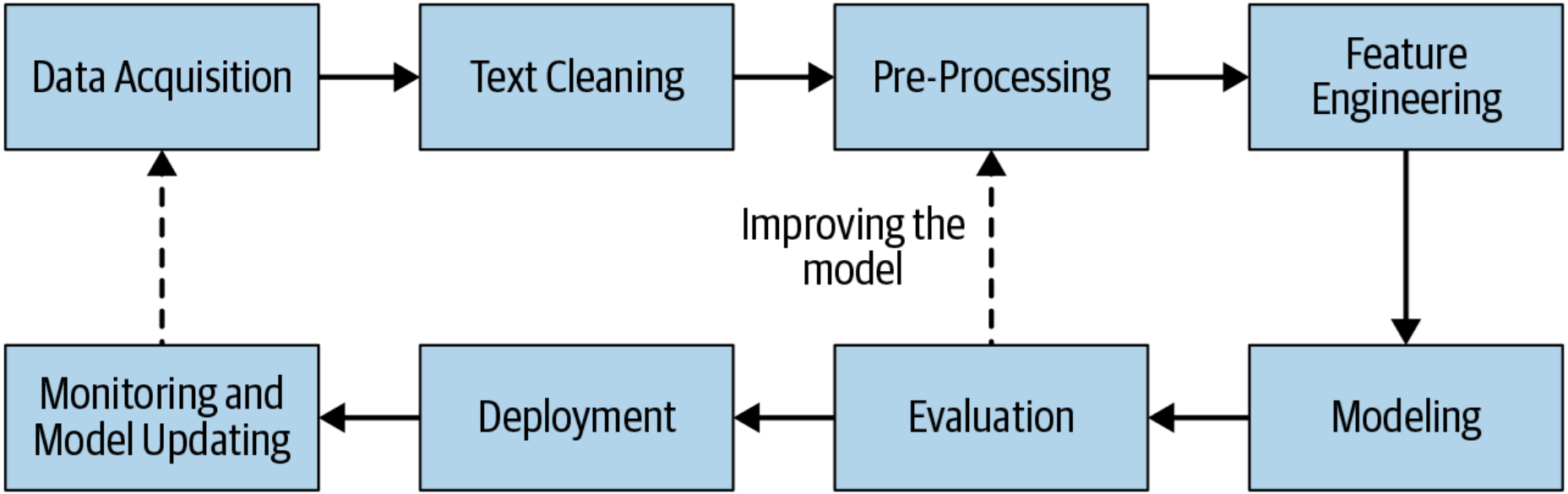
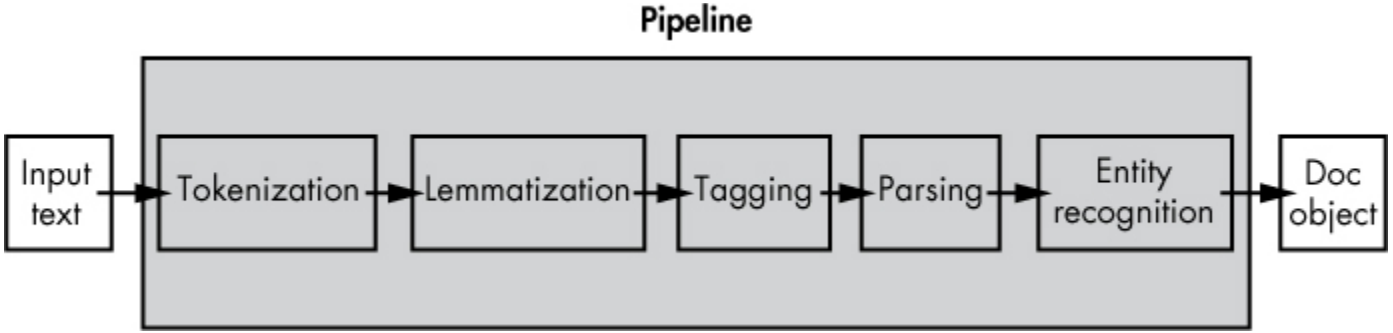
Técnicas de Procesamiento del Habla

Spacy

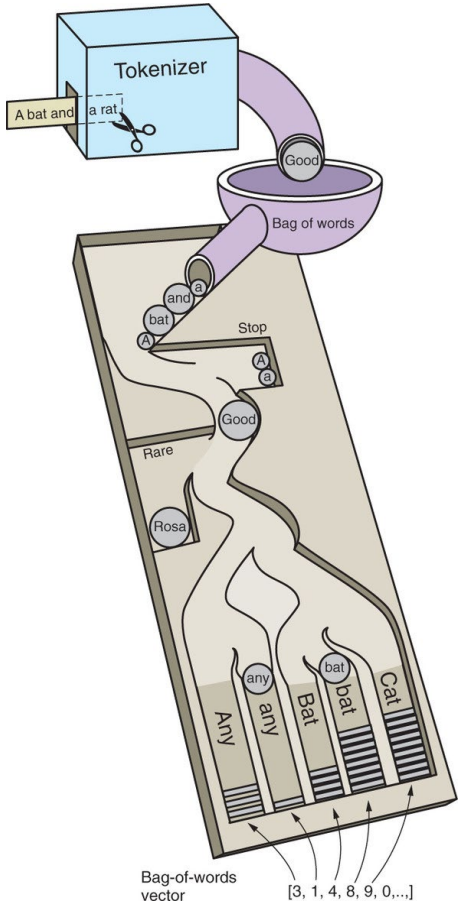

Natural language understanding applications
Interactive
Intelligent assistants
Education
Interactive translation
Tutoring
Generic voice assistants
Enterprise assistants
Language learning
Chatbots
Voice based


Natural language understanding applications
Non-Interactive
Classification
Translation
Document retrieval
Simplification
Summarization
Information extraction
Document classification
Sentiment analysis
Detecting malicious documents
Fake news detection
Phishing detection
Spam detection


Data Acquisition
Text Cleaning
Pre-Processing
Feature Engineering
Improving the model
Monitoring and Model Updating
Deployment
Evaluation
Modeling


Pipeline
Input text
Tokenization
Lemmatization
Tagging
Parsing
Entity recognition
Doc object

Tokenización: Es el primer paso crucial en el procesamiento de texto. **Consiste en dividir el texto de entrada en unidades individuales llamadas tokens. Estos tokens representan las piezas fundamentales del texto, como palabras, números y signos de puntuación.** La tokenización es esencial porque sirve como base para todas las operaciones posteriores de análisis lingüístico. Por ejemplo, *la oración "El perro ladra fuerte." se tokenizaría en: ['El', 'perro', 'ladra', 'fuerte', '.']*



```
nlp = spacy.load('es_core_news_sm')
doc = nlp(u'Estoy volando hacia Buenos Aires')
print([w.text for w in doc])
```

Lematización: Es el proceso de **reducir una palabra a su forma base o de diccionario, conocida como lema.** Esto implica identificar la raíz de la palabra, eliminando sufijos flexionales (como terminaciones de plural o conjugaciones verbales). Por ejemplo, el lema de "corriendo" es "correr". **La lematización es importante para el reconocimiento de significado, ya que agrupa diferentes formas de la misma palabra bajo una única representación**, simplificando el análisis

```
doc = nlp(u'En abril de 2025 estaremos realizando una introducción a spaCy para trabajar conceptos basicos de PNL')
for token in doc:
  print(token.text, token.lemma_)
```

Etiquetado de Partes del Discurso (Part-of-Speech Tagging): Consiste en asignar una etiqueta gramatical a cada token en el texto, indicando su función sintáctica (por ejemplo, sustantivo, verbo, adjetivo, adverbio). Estas etiquetas pueden ser coarse-grained (generales) o fine-grained (específicas). **El etiquetado POS es fundamental para comprender la estructura gramatical de las oraciones y para tareas como identificar verbos relevantes para la extracción de intenciones.**

```
from spacy.lang.es.examples import sentences

doc = nlp(sentences[1])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

Análisis de Dependencias Sintácticas (Syntactic Dependency Parsing): Este proceso analiza la estructura gramatical de una oración, identificando la palabra raíz (head) y las relaciones de dependencia sintáctica entre las palabras (dependents). **Estas relaciones se visualizan a menudo como un árbol de dependencias, donde cada nodo es un token y las aristas representan las dependencias.** El parsing de dependencias es crucial para entender cómo las palabras se relacionan entre sí gramaticalmente y para extraer el significado y las relaciones semánticas dentro de las oraciones.

Reconocimiento de entidades nombradas (Named Entity Recognition - NER): Este proceso identifica y clasifica menciones de entidades del mundo real que aparecen en el texto. Estas entidades pueden pertenecer a categorías como personas, organizaciones, lugares (GPE - geopolitical entity, como BA o Baires), fechas, cantidades de dinero, etc.. El NER es fundamental para extraer información específica del texto.