NLP

Subash Gandyer
Data Scientist, HealthChain

Introduction to NLP

NLP

- What is NLP?
- Applications
- NLP Approaches

What is NLP?

Applications

NLP Applications

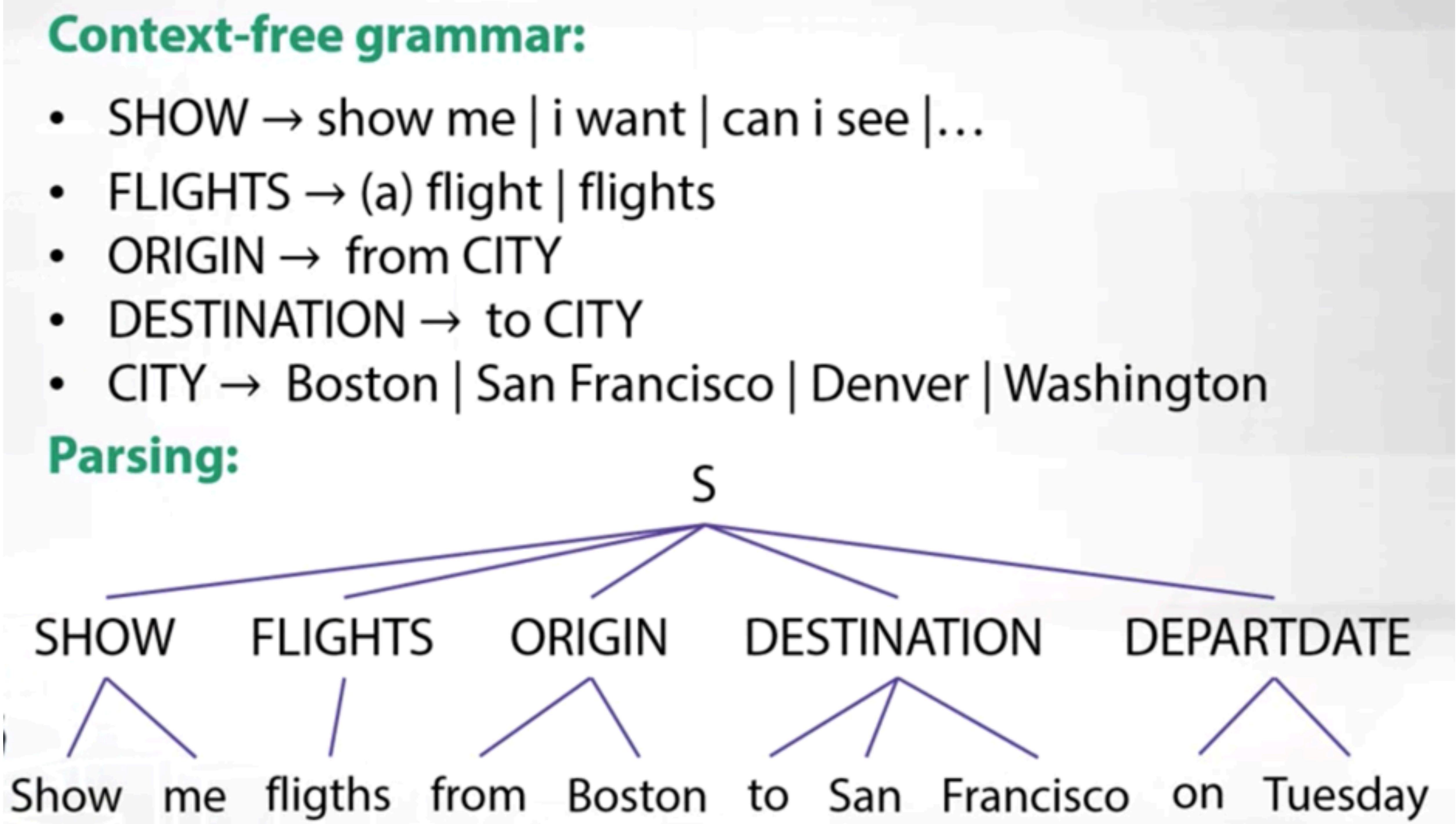
- Part-of-speech tagging: identify if each word is a noun, verb, adjective, etc.)
- Named entity recognition NER): identify person names, organizations, locations, medical codes, time expressions, quantities, monetary values, etc)
- Question answering
- Text Summarization
- Text-to-speech and Speech-to-text
- Topic modeling
- Sentiment classification
- Language modeling
- Translation

NLP Approaches

Approaches

- Rule Based
  - Regular Expressions
  - Context-free Grammars
- Machine Learning
  - Probabilistic Modeling
  - Linear Classifiers
- Deep Learning
  - Recurrent Neural Networks
  - Convolutional Neural Networks

Semantic Slot Filling



Context-free grammar:

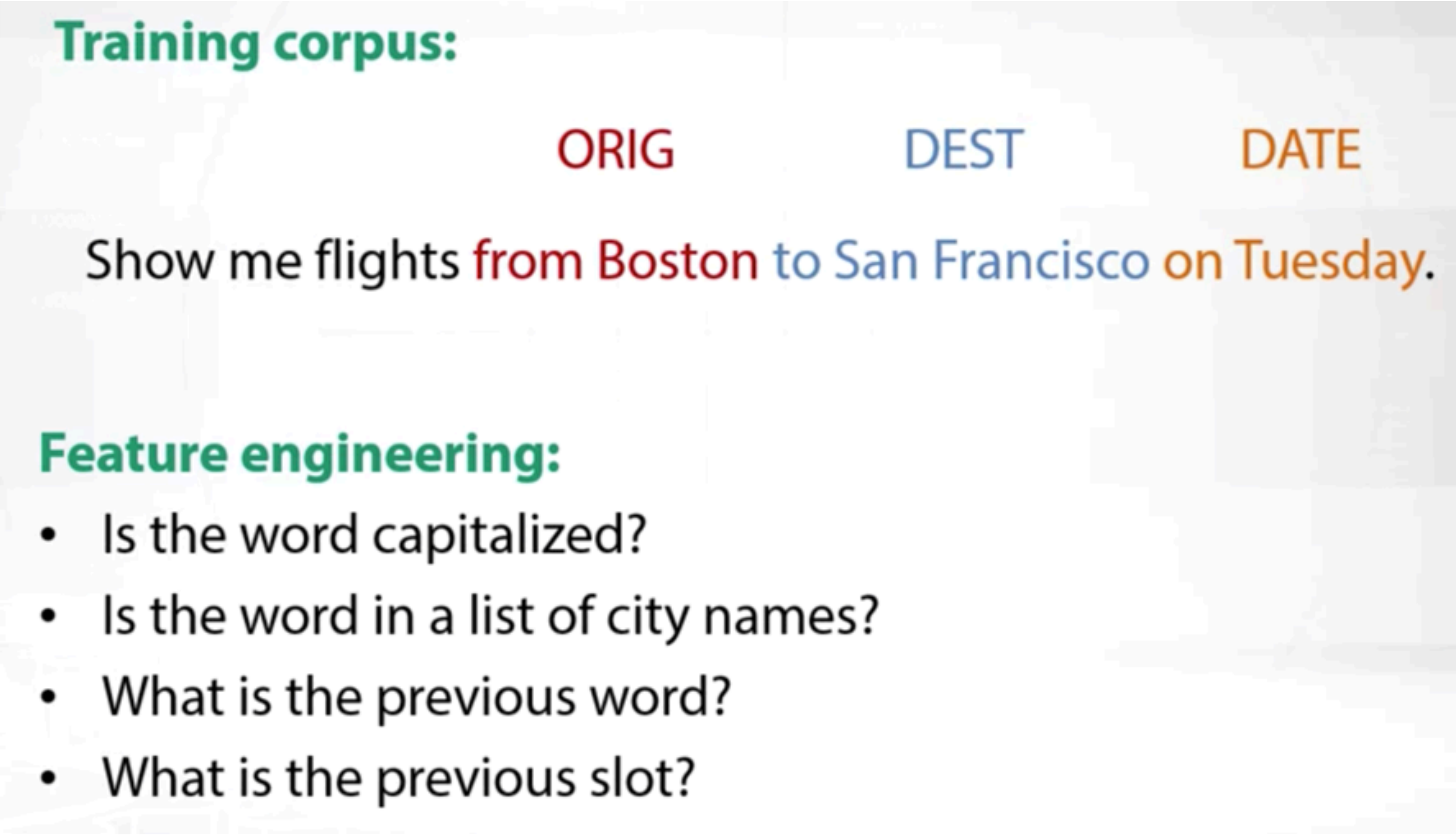
- SHOW → show me | i want | can i see |…
- FLIGHTS → (a) flight | flights
- ORIGIN → from CITY
- DESTINATION → to CITY
- CITY → Boston | San Francisco | Denver | Washington

Parsing:

S

SHOW FLIGHTS ORIGIN DESTINATION DEPARTDATE

Show me fligths from Boston to San Francisco on Tuesday



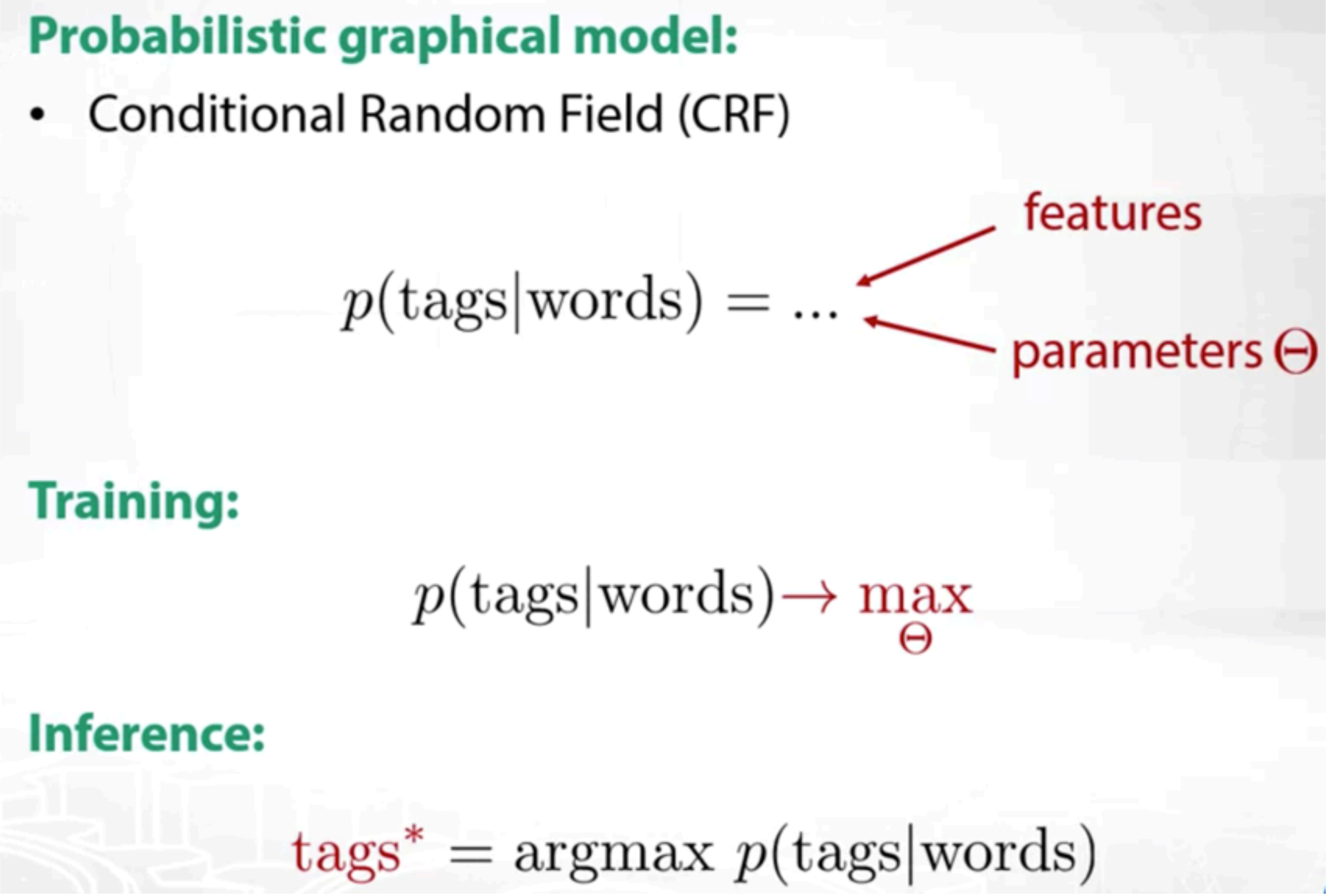
Training corpus:

ORIG DEST DATE

Show me flights from Boston to San Francisco on Tuesday.

Feature engineering:

- Is the word capitalized?
- Is the word in a list of city names?
- What is the previous word?
- What is the previous slot?



Probabilistic graphical model:

- Conditional Random Field (CRF)

$$p(\text{tags}|\text{words}) = \ldots$$

features

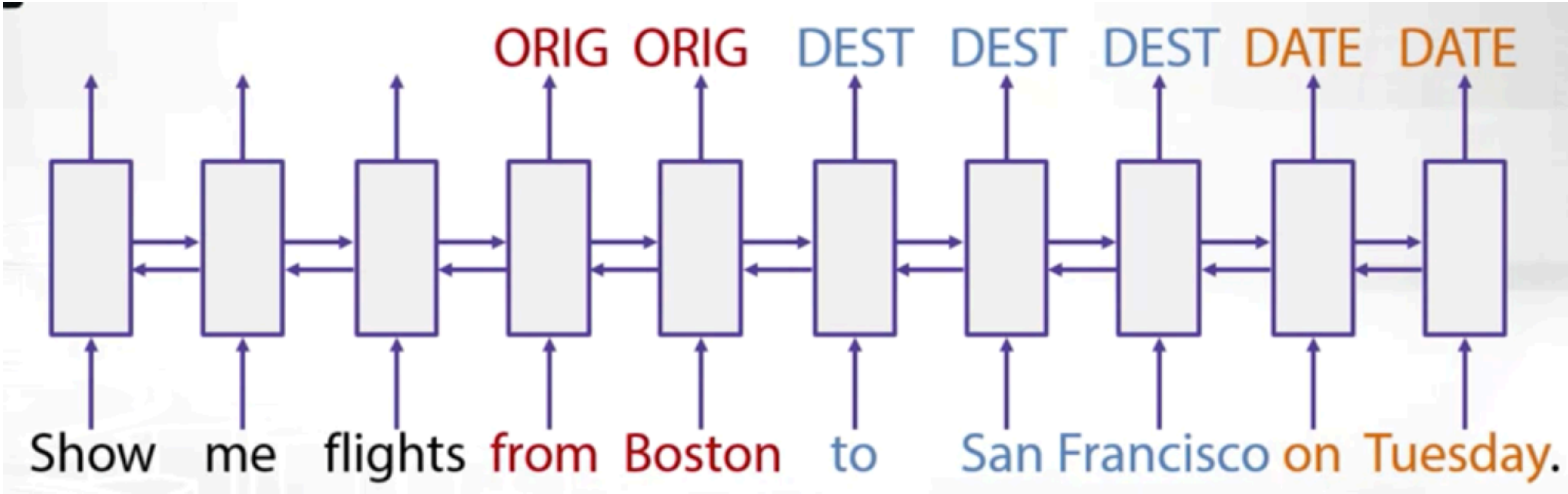
parameters Θ

Training:

$$p(\text{tags}|\text{words}) \rightarrow \max_{\Theta}$$

Inference:

$$\text{tags}^* = \text{argmax}\ p(\text{tags}|\text{words})$$



ORIG ORIG DEST DEST DEST DATE DATE
Show me flights from Boston to San Francisco on Tuesday.

Fundamentals of Text Processing

Text Classification

- Predict Tags or Categories
- Predict Sentiment
- Filter Spam mails

Sequence applications

- Part Of Speech Tags
- Named Entity Recognition
- Semantic Slots

Representations

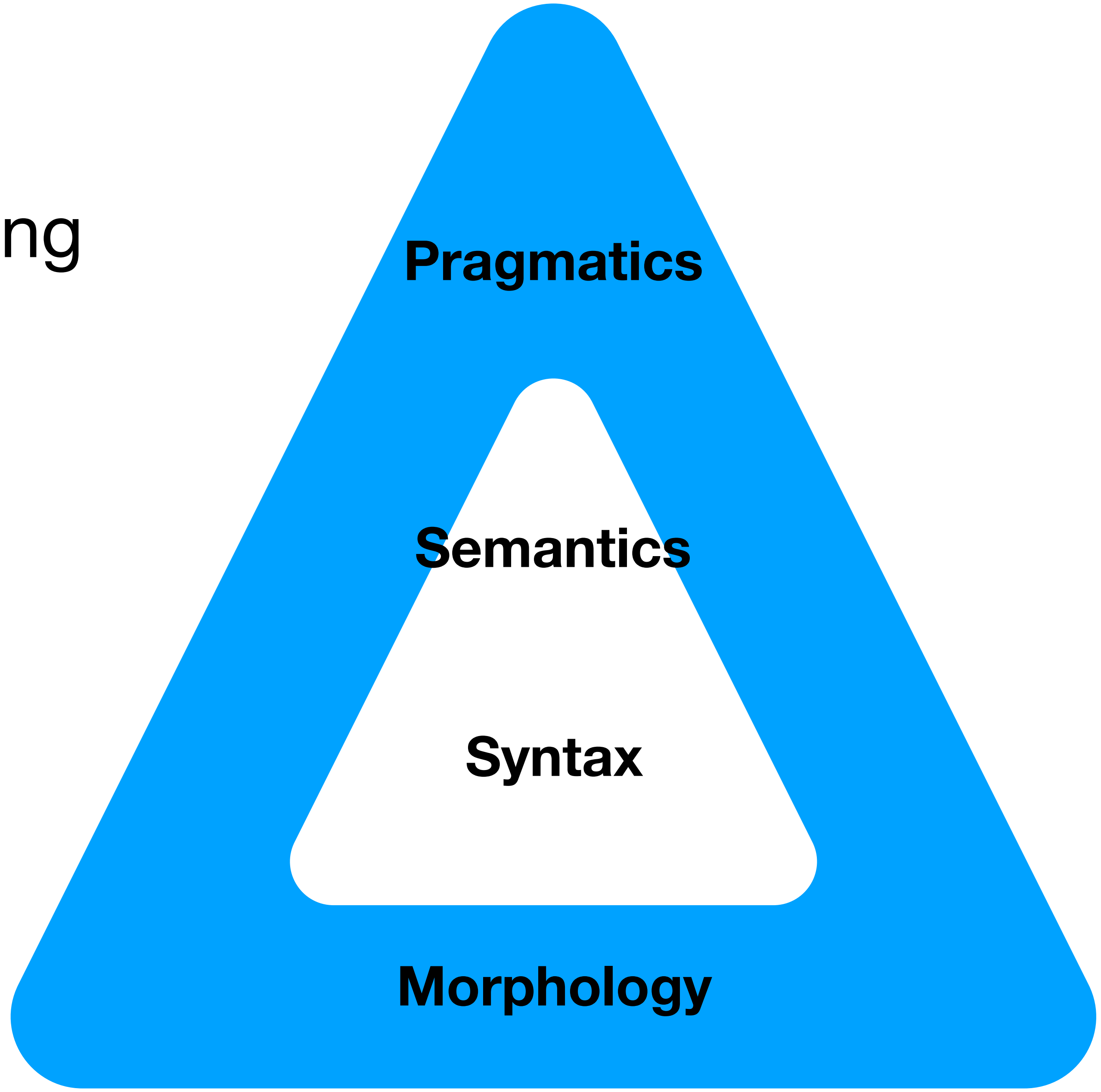
- Word Embeddings
- Sentence Embeddings
- Topic Models (Documents)
- Vector Space Models
- Similarity Graphs

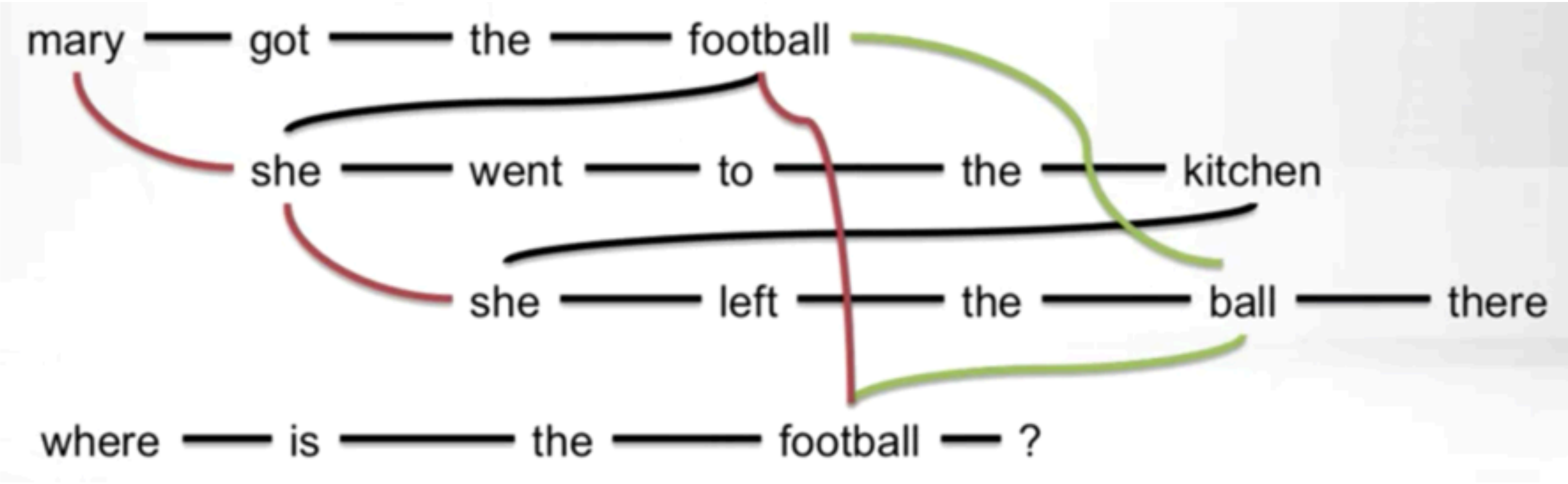
Sequence to Sequence

- Machine Translation
- Summarization
- Speech Recognition
- Question Answering

Linguistic Pyramid

- Morphology - Pre-processing
- Syntax
- Semantics
- Pragmatics





mary got the football

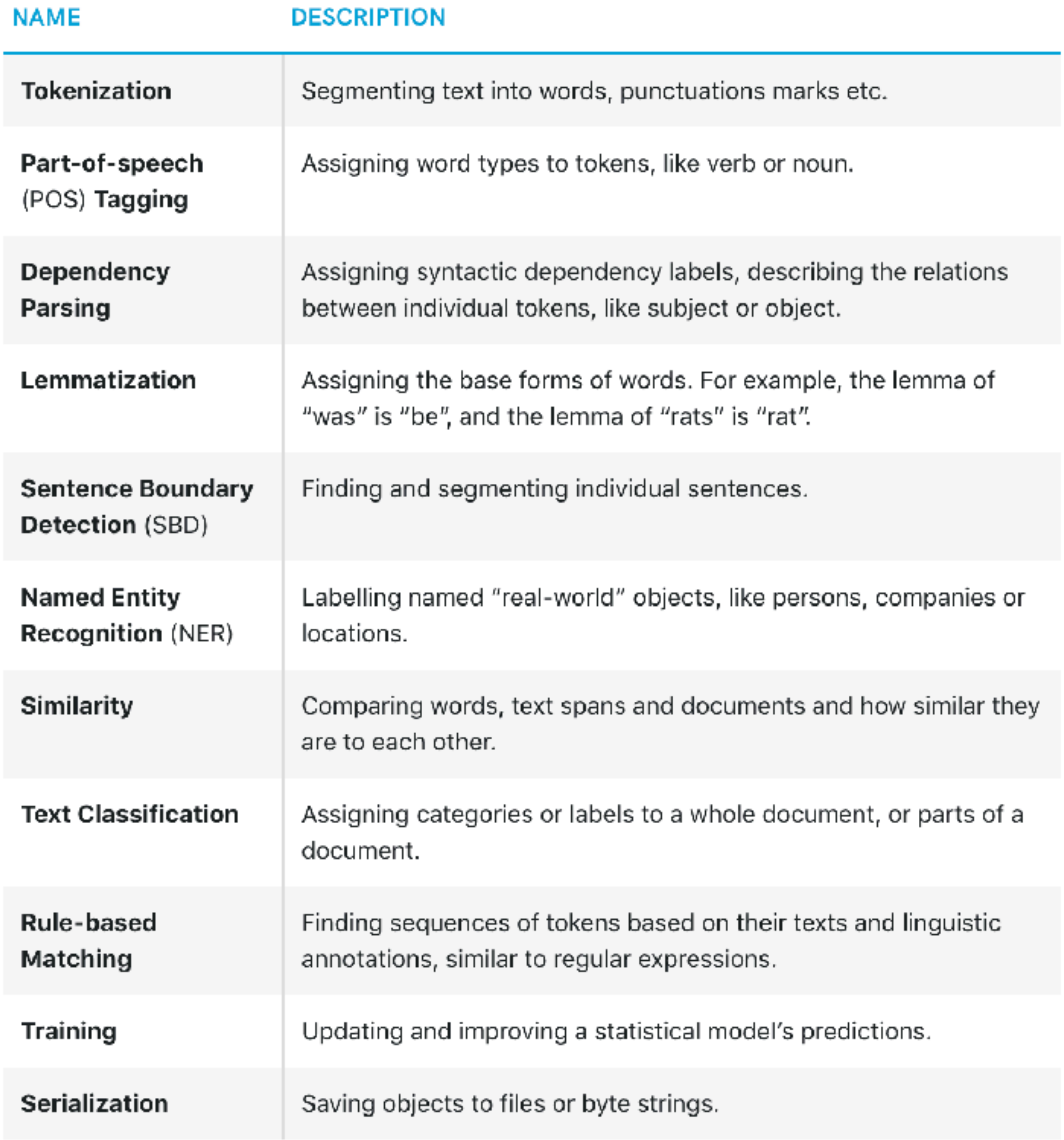
she went to the kitchen

she left the ball there

where is the football ?

Pre-processing

- Tokenization
- Parts Of Speech Tagging
- Stemming
- Lemmatization



NAME	DESCRIPTION
Tokenization	Segmenting text into words, punctuations marks etc.
Part-of-speech (POS) **Tagging**	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, like subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of “was” is ”be”, and the lemma of “rats” is “rat”.
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named “real-world” objects, like persons, companies or locations.
Similarity	Comparing words, text spans and documents and how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document, or parts of a document.
Rule-based Matching	Finding sequences of tokens based on their texts and linguistic annotations, similar to regular expressions.
Training	Updating and improving a statistical model’s predictions.
Serialization	Saving objects to files or byte strings.

NLP Frameworks

Popular NLP Frameworks

- Stanford Core NLP Parser
- TextBlob
- Flair
- Spacy
- NLTK

Stanford NLP

01-StanfordNLP

TextBlob

02-TextBlob

Flair

03-Flair

Spacy

NLTK

Feature Extraction

- document - refers to a single piece of text information. This could be a text message, tweet, email, book, lyrics to a song. This is equivalent to one row or observation.
- corpus - a collection of documents. This would be equivalent to a whole data set of rows/observations.
- token - this is a word, phrase, or symbols derived from a document through the process of tokenization. This will happen behind the scenes so we won’t need to worry too much about it and for our purposes it essentially means a word. For example the document `'How are you’` would have tokens of `'How'`, `'are'`, and `‘you'`

```
messages = ["Hey hey hey lets go get lunch today :)",
            "Did you go home?",
            "Hey!!! I need a favor"]
```

Bag Of Words

Theory

Suppose we have a corpus with three sentences:

- "I like to play football"
- "Did you go outside to play tennis"
- "John and I play tennis”

Goal: Convert text to numbers

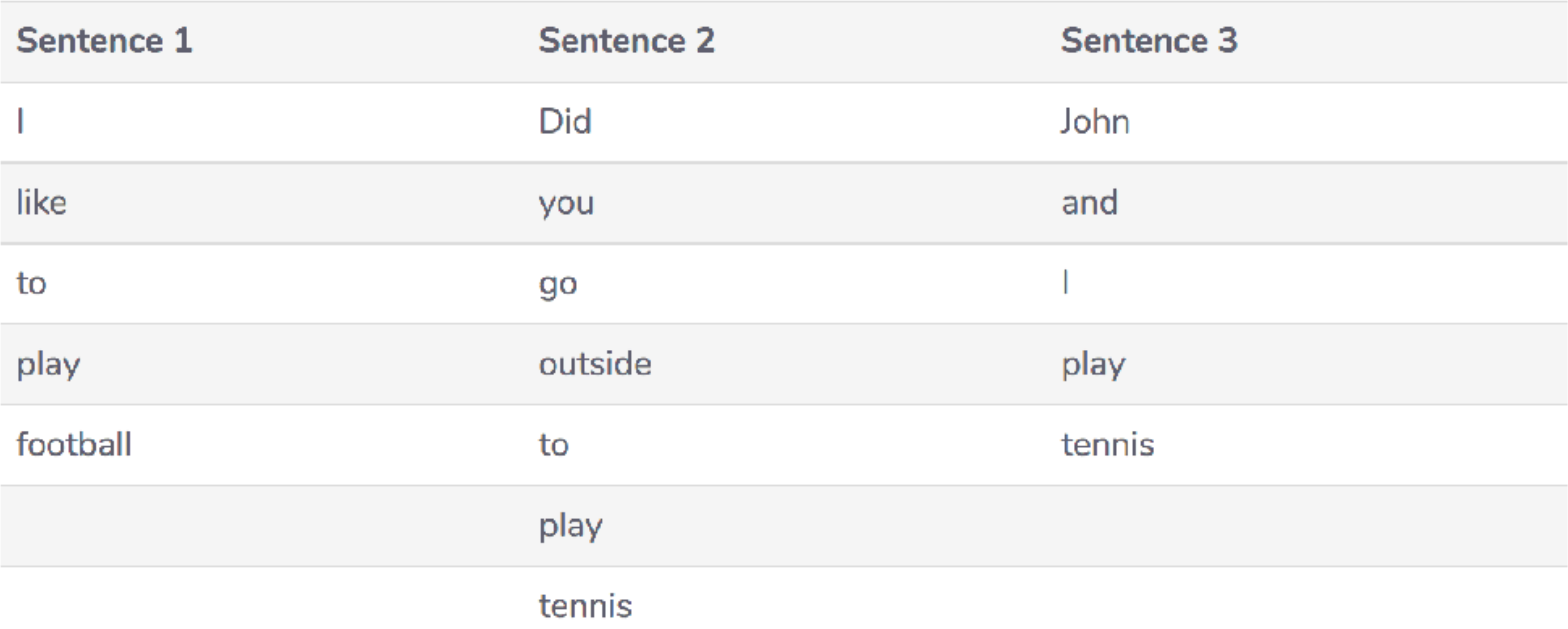
Steps

1. Tokenize the sentences into words

2. Create Dictionary of Word Frequency

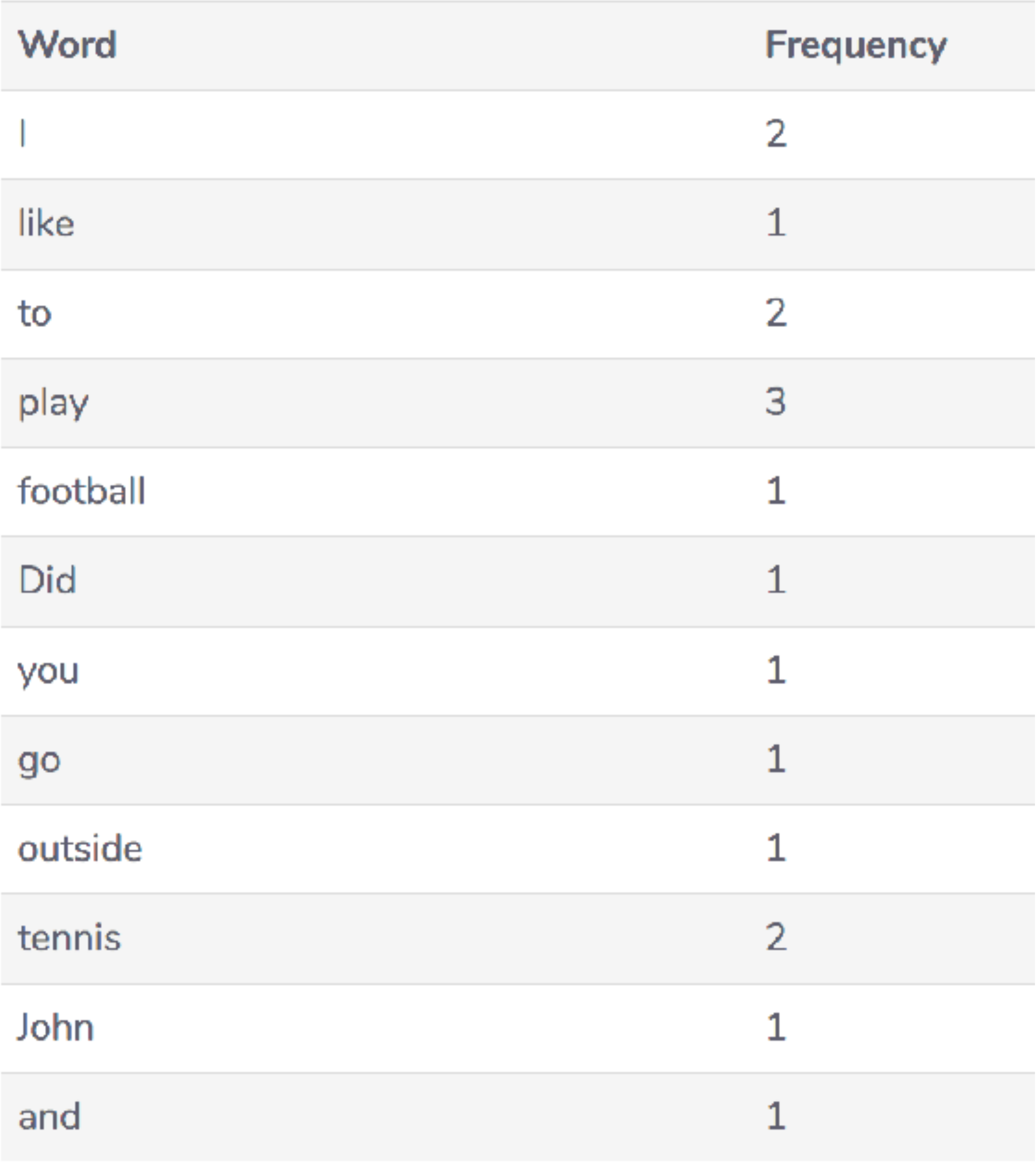
3. Bag of Words Model

Step 1: Tokenization



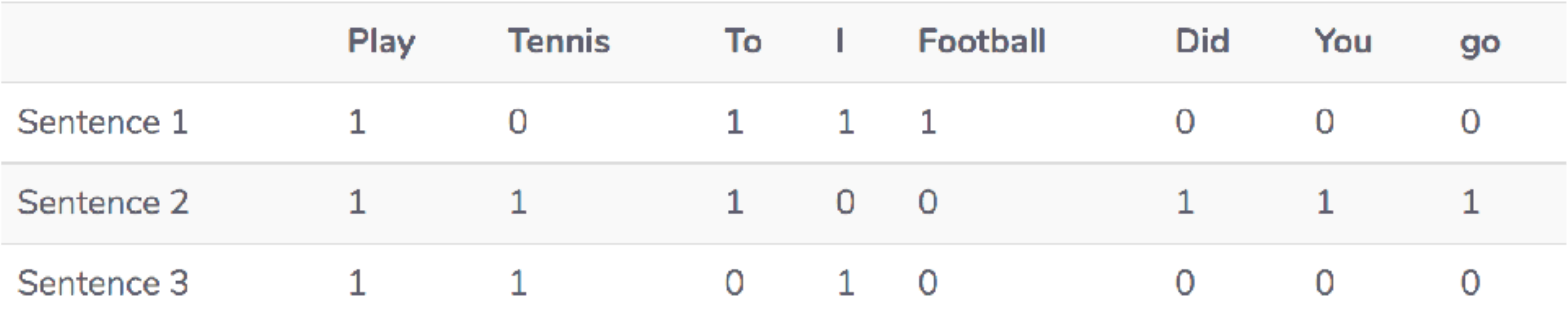
Sentence 1	Sentence 2	Sentence 3
I	Did	John
like	you	and
to	go	I
play	outside	play
football	to	tennis
	play	
	tennis	

Step 2: Dictionary of word frequency



Word	Frequency
I	2
like	1
to	2
play	3
football	1
Did	1
you	1
go	1
outside	1
tennis	2
John	1
and	1

Step 3: Bag of Words Model



	Play	Tennis	To	I	Football	Did	You	go
Sentence 1	1	0	1	1	1	0	0	0
Sentence 2	1	1	1	0	0	1	1	1
Sentence 3	1	1	0	1	0	0	0	0

Let’s Create a Bag Of Words Model

CountVectorizer

TF-IDF Vectorizer

NLP Advancements
SOTA Algorithms