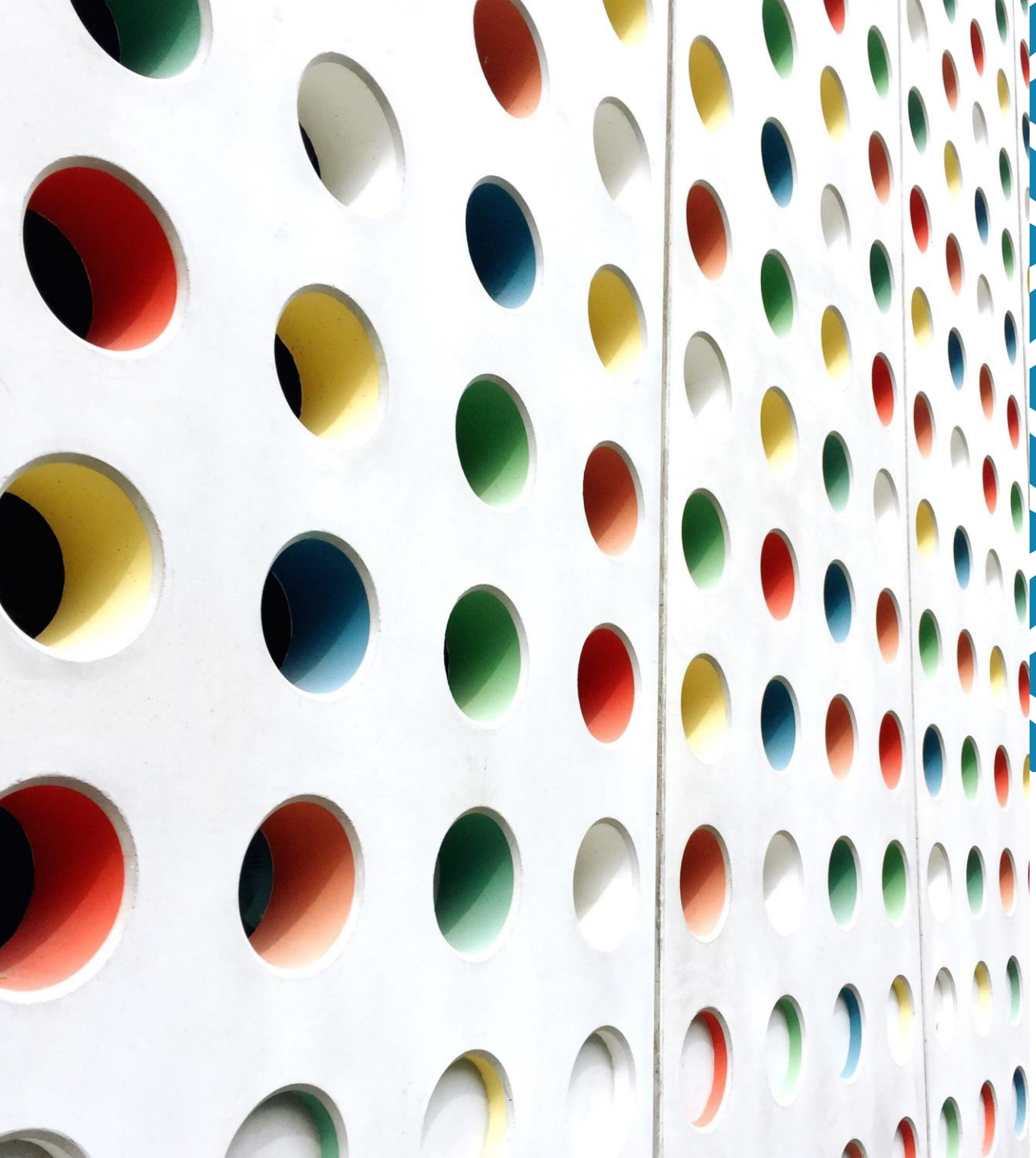


NLP- MODULE 6

Hazina
DCS KU

NAMED ENTITY RECOGNITION AND RELATION EXTRACTION

- An information extraction technique that identifies key elements from text then classifies them into predefined categories.
- Text appears as unstructured data in several formats such as document files, spreadsheets, web pages and social media.
- Need for NER!
- With hundreds of documents to review, or to investigate, its hard to obtain the content for related purpose.
- So,-Entity extraction is performed.

ENTITY EXTRACTION

- Entity extraction can provide a useful view of unknown data sets by immediately revealing who and what the information at a minimum.
- In a structured corpus that they can use as a point of departure for further analysis and investigation.
- Must address a number of language issues to correctly identify and classify entities.
- The ambiguities of language make this an especially complex task for machines.
- Part Of Speech Tagging (main challenge)
- EE-This function reveals direct relationships, connections or events shared between different entities as well as complex relationships through inferred, indirect connections. This helps to better summarize information in a quick and efficient manner.

RELATION EXTRACTION

Extracting relations among named entities using NLP.

Consider these sentences.

Maya Mytri works at Infosys.
Apple is located in Cupertino.

Detect the various entities in these sentences.

Person name = Maya Mytri
Company = Infosys
Company = AppleCity = Cupertino

Deduce the relationships using NLP

works_at(person: Maya Mytri,company:infosys)
located_in(company:apple, city:cupertino)

CONTD.

POS tags carry strong signals towards identifying the relational phrases.

Also *named entity recognition* would have already tagged the named entities in these sentences which would additionally make easier.

Tokens: Maya Mytri works at Infosys
POS tags: noun noun verb preposition noun
NER tags: person person — — companyTokens: Apple is located in Cupertino
POS tags: noun verb verb preposition noun
NER tags: company — — — city

combining the information in the POS tags with those in the NER tags works well to identify the relational phrases.

CONTD.

Consider this sentence,

Read The Adventures of Sherlock Holmes by Arthur Conan Doyle online or in your email

Named entities are *The Adventures of Sherlock Holmes, a book title, and Arthur Conan Doyle, a person's name*.

(Relational phrase still sits between them after recognizing)

-This sentence has the structure;

<words> book-title by person-name <words>

POS analysis= The Adventures of Sherlock Holmes by Arthur Conan Doyle

NER= Able to infer that the book title always precedes the *by* and that the person's name always follows it

CONTD.

Wimbledon is a tennis tournament held in the UK in the first two weeks of July every year.

could be ‘parsed’ via entity recognition*;*

Sports event name is a **sport tournament** held in the **location** in the **time**.

phases of recognition

Sports event name is-a **sport tournament*.=it is crisp and hence get easily trained and tested with maximum accuracy, but held in may be wrong inference so, we have inferences,***

ST=tennis tournament, held-in ⇒ L=UK
SEN=Wimbledon, held-in ⇒ L=UK
SEN=Wimbledon, ST=tennis tournament, held-in ⇒ L=UK

MATHEMATICAL INFERENCE/GENERALIZATION

IF X THEN Y;

he support of a rule $X \Rightarrow Y$ is the number of sentences in which X holds. The confidence of this rule is $P(Y|X)$.

In $X \Rightarrow Y$, Y is always **T** = *first two weeks of July every year*. **X** is any nonempty subset of {**SEN**=*Wimbledon*, **ST**=*tennis tournament*, **L**=*UK*}

INFORMATION EXTRACTION USING SEQUENCE LABELING

Information Extraction is the process of parsing through unstructured data and extracting essential information into more editable and structured data formats.

Information Extraction NLP techniques,

1. Tokenization
2. **Parts of Speech Tagging**
3. **Dependency Graphs**
4. **NER with Spacy**

SEQUENCE LABELING

- **Sequence labeling** is a fundamental technique in NLP that is used to *identify* and *label* the components of a sequence, such as words or phrases in a sentence.
- A preprocessing step for other NLP tasks
- In **Information Retrieval**, it helps to clarify the context and meaning of a query. Additionally, sequence labeling is employed in **machine translation** to identify the grammatical structure of a sentence and to facilitate the translation process.
- POS and Named Entities are useful clues to sentence structure and meaning
- POS tagging- taking a sequence of words and assigning each word a part of speech like NOUN or VERB
- NER- assigning words or phrases tags like PERSON, LOCATION, or ORGANIZATION.

CONTD,.

- Each word x_i in an input word sequence, a label y_i , so that the output sequence Y has the same length as the input sequence X are called **Sequence Labeling Tasks**

APPROACHES TO SEQUENCE LABELING TASKS

There are multiple ways to perform those tasks, and the method chosen can significantly impact the performance and outcome.

- **Rule-based approaches** : These rely on a set of manually-defined rules going from predefined rules to tag each word in a sentence or identifying named entities in text. These do the job for simple tasks but can be error-prone and time consuming.
- **Machine learning-based approaches** : These approaches use machine learning techniques to learn the patterns for the given tasks from annotated training data. They range from **Stochastic approaches** to **Deep learning-based approaches** such as Transformers.
- **Hybrid approaches** : These approaches combine the strengths of rule-based and statistical approaches, using a combination of hand-written rules and machine learning techniques to identify arguments and roles.

TYPE OF SEQUENCE LABELLING

1. **Part of Speech Tagging**-**Part of speech tagging problem generate different**

   Part of speech tag from an input sentence

   For example: *I eat rice -> PRON VERB NOUN*

2. **Lemmatization**-Lemmatization predict different lemma from an input sentence (reduce a word to its root form/grouping together different inflected forms of the same word)

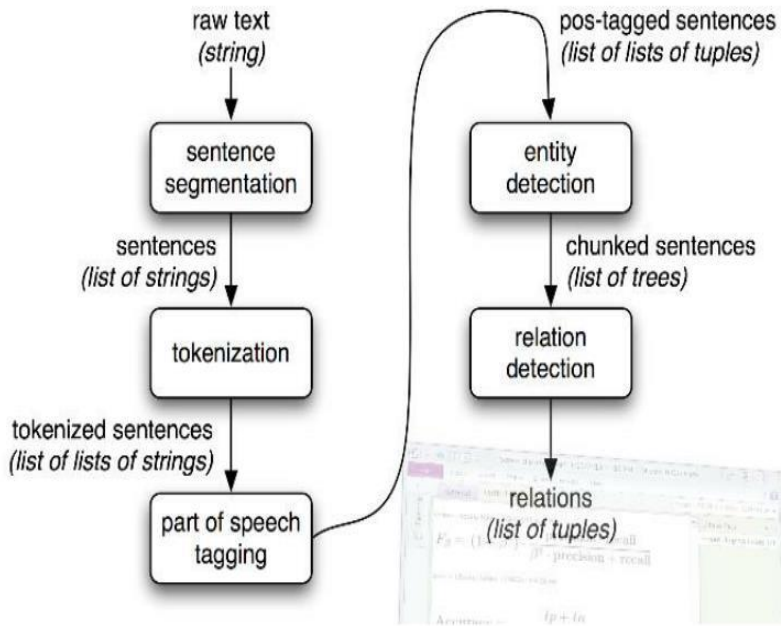
   He ate Rice-> He eat Rice

3. **Language Identification**

   Language identification predict different language type from an input sentence

   el comio arroz-Spanish

4. **NER**


raw text
(string)
sentence segmentation
sentences
(list of strings)
tokenization
tokenized sentences
(list of lists of strings)
part of speech tagging
pos-tagged sentences
(list of lists of tuples)
entity detection
chunked sentences
(list of trees)
relation detection
relations
(list of tuples)

EXAMPLE

CV: Data scientist, hands-on expertise in machine learning, big data, development, statistics, and analytics. With my team of data, scientists implemented Python machine learning model ensembles, stacking, and feature engineering demonstrating high accuracy rates in predictive analytics. Created a recommender system using Doc2Vec words embeddings and neural networks

Extracted professional skills: machine learning, big data, development, statistics, analytics, Python machine learning model ensembles, stacking, feature engineering, predictive analytics, Doc2Vec, word embeddings, neural networks.

5 CONDITIONAL RANDOM FIELDS (CRFS)

MACHINE TRANSLATION IN NLP?

VITERBI