

# Social Media Analytics (SMA)

## *Named Entity Recognition, Linking, and Disambiguation*

### *Part 1*

**Marco Viviani**

University of Milano-Bicocca

*Department of Informatics, Systems, and Communication*



DIPARTIMENTO DI  
INFORMATICA, SISTEMISTICA E  
COMUNICAZIONE

# Ambiguity in NLP (1)

- **Emotional ambiguity**
  - Sentiment analysis
  - Irony detection

New Frozen Boutique to Open at Disney's Hollywood Studios.  
Can't wait 😞

# Ambiguity in NLP (2)

- **Emotional ambiguity**
  - Sentiment analysis
  - Irony detection

New Frozen Boutique to Open at Disney's Hollywood Studios.

Can't wait 😞

# Ambiguity in NLP (3)

- **Semantic ambiguity**

- Named-entity Recognition (NER)
- Named-entity Linking/Disambiguation (NEL/NED)

New Frozen Boutique to Open at Disney's Hollywood Studios.

Can't wait 😞

# Ambiguity in NLP (4)

- **Semantic ambiguity**

- Named-entity Recognition
- Named-entity Linking/Disambiguation (NEL/NED)

New Frozen Boutique to Open at Disney's Hollywood Studios.

Can't wait 😞

[https://dbpedia.org/page/Frozen\\_\(2013\\_film\)](https://dbpedia.org/page/Frozen_(2013_film))



[https://dbpedia.org/page/Frozen\\_\(Madonna\\_song\)](https://dbpedia.org/page/Frozen_(Madonna_song))

# Ambiguity in NLP (5)

- **Semantic ambiguity**

- Named-entity Recognition
- Named-entity Linking/Disambiguation (NEL/NED)

New Frozen Boutique to Open at Disney's Hollywood Studios.

Can't wait 😞

[https://dbpedia.org/page/Frozen\\_\(2013\\_film\)](https://dbpedia.org/page/Frozen_(2013_film))



[https://dbpedia.org/page/Frozen\\_\(Madonna\\_song\)](https://dbpedia.org/page/Frozen_(Madonna_song))

[https://dbpedia.org/page/The\\_Walt\\_Disney\\_Company](https://dbpedia.org/page/The_Walt_Disney_Company)



# Ambiguity in NLP (6)

- **Semantic ambiguity**

- Named-entity Recognition
- Named-entity Linking/Disambiguation (NEL/NED)

New Frozen Boutique to Open at Disney's Hollywood Studios.

Can't wait 😞

[https://dbpedia.org/page/Frozen\\_\(2013\\_film\)](https://dbpedia.org/page/Frozen_(2013_film))



[https://dbpedia.org/page/Frozen\\_\(Madonna\\_song\)](https://dbpedia.org/page/Frozen_(Madonna_song))

[https://dbpedia.org/page/Disney's Hollywood Studios](https://dbpedia.org/page/Disney's_Hollywood_Studios)



# NER, NEL, and NED (1)

- **Named Entity Recognition (NER)**

- NER is a task that involves **identifying** and **classifying** entities, such as names of people, organizations, locations, dates, and other predefined categories, within a given text.
- The primary goal of NER is to provide **structured information** about the entities present in a text.



# NER, NEL, and NED (2)

- **Named Entity Linking (NEL)**

- NEL involves **linking entities** to a specific **entry** or **entity** in a **knowledge base** or **reference database**.
- The goal is to associate each recognized entity in the text with a **unique identifier** in a knowledge base (or reference database), allowing for additional information retrieval or disambiguation.
  - One **term** can be associated with **one or more entities** (entities are **ambiguous**).

# NER, NEL, and NED (3)

- **Named Entity Disambiguation (NED)**

- NED is the task of determining the specific entity (e.g., person, location, organization) that a mention or reference in the text corresponds to.
- This task is particularly important in cases where the same term could refer to multiple entities.
- Disambiguation helps to identify the correct entity based on the context.

# NER, NEL, and NED (4)

- In summary, **NER** involves identifying and classifying entities in a text, while **NED** deals with resolving ambiguity when a name or term could refer to multiple entities by determining the correct entity based on context.
- While **NEL** associates the recognized entity with a unique identifier in a knowledge base, **NED** focuses on resolving ambiguity by determining the correct entity in a given context.

# Challenges (1)

- NER, NEL, and NED in social media texts (e.g., Tweets) has been reported to be **challenging**:
  - Short and noisy nature, typographic errors, shortening of words, ambiguity, polysemy
    - Derczynski, Leon, et al. "[Analysis of named entity recognition and linking for tweets.](#)" *Information Processing & Management* 51.2 (2015): 32-49.
    - Carmel, David, et al. "[ERD'14: entity recognition and disambiguation challenge.](#)" *Acm Sigir Forum*. Vol. 48. No. 2. New York, NY, USA: Acm, 2014.
    - Goyal, Archana, Vishal Gupta, and Manish Kumar. "[Recent named entity recognition and classification techniques: a systematic review.](#)" *Computer Science Review* 29 (2018): 21-43.
  - ...

# Challenges (2)

- ...
- **Out of Vocabulary (OOV)** entity mention identification problem.
  - The Big Bang Theory being referred as TBBT.
- **Out of Knowledge Base (OOKB)** entity problem.
  - A new upcoming company *Widro*.
- **Named Entity overlap**
  - Social media posts often contain overlapping named entities, such as hashtags.
- **User-generated Entities**
  - Users on social media platforms often create new terms, hashtags, or nicknames for entities.
- **Context variability**
  - Entities can change roles and relationships rapidly on social media.
  - The same entity may be mentioned in various contexts, and discerning the correct context is crucial for accurate disambiguation.

# Named Entity Recognition

---

# Deepening NER

- **Named Entity Recognition (NER)** is an **Information Extraction (IE)** task that seeks to segment and classify text fragments into predefined classes/labels (e.g., `Person`, `Location`, `Organization`, ...).
  - IE is an NLP task that involves automatically extracting **structured information from unstructured text**.
    - The goal is to transform raw text into a more organized and structured format, making it easier to analyze and use.
  - Information extraction typically involves **identifying** and **extracting specific types of information**, such as entities, relationships, and events, from text documents.

# NER: An example

CRICKET – **MILLNS** SIGNS FOR  
**BOLAND**

**CAPE TOWN** 1996-08-22

South African provincial side **Boland** said on Thursday they had signed **Leicestershire** fast bowler **David Millns** on a one year contract. **Millns**, who toured **Australia** with **England** A in 1992, replaces former **England** all-rounder **Phillip DeFreitas** as **Boland's** overseas professional.

Labels:

Examples:

**PER**

David Millns  
Philip DeFreitas

**ORG**

Boland

**LOC**

Cape Town  
Australia  
England



# Sequence Prediction Problem

- The **Sequence Prediction Problem (SPP)** refers to a type of machine learning task where the goal is to predict the next element in a sequence based on the patterns and relationships observed in the input data.
- In this context, a "sequence" refers to an ordered set of data points, and the objective is to model the underlying patterns in the sequence to make predictions about what comes next.

# SPP: Main characteristics

- **Sequential data**

- The input data consists of **sequences**, where each element is **ordered** and has a **temporal** or **spatial relationship** with the preceding and succeeding elements.

- **Prediction task**

- The primary objective is to **predict the next element** in the sequence.
- The prediction could be based on various factors, such as **historical information**, **contextual dependencies**, or **patterns** learned from the data.

- **Types of sequences**

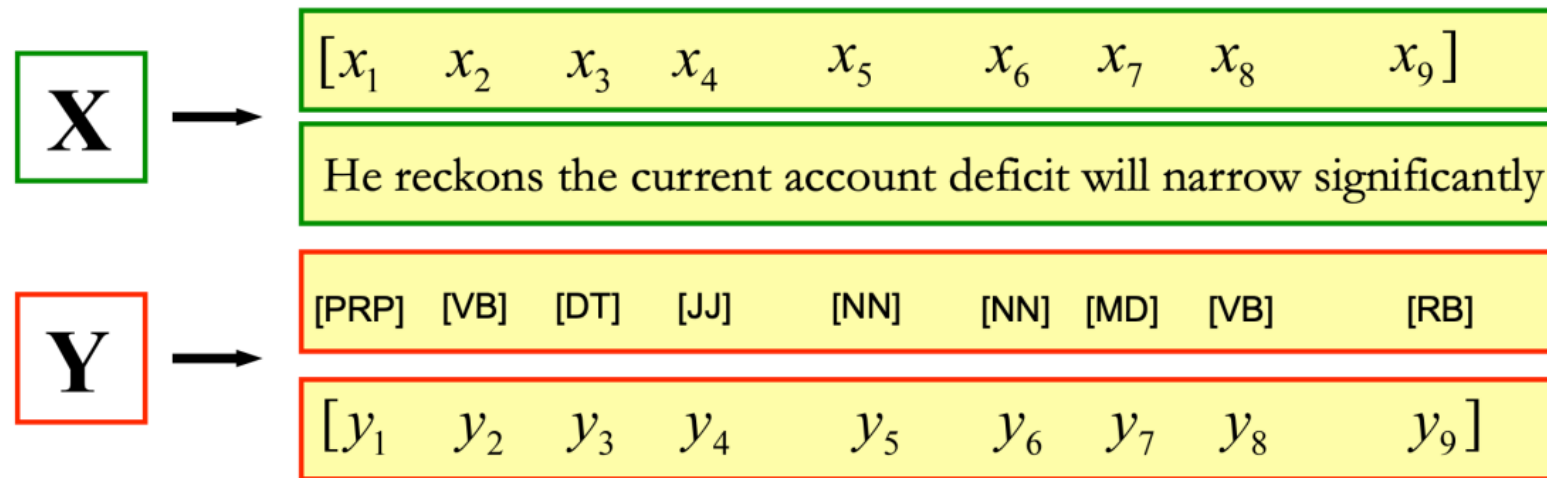
- Sequences can take different forms, such as **text sentences**, **time series data** (e.g., stock prices, weather measurements), **DNA sequences**, and more.

- **Algorithms**

- **Various machine learning algorithms** can be applied to address sequence prediction problems.
- **Recurrent Neural Networks (RNNs)**, **Long Short-Term Memory networks (LSTMs)**, and **Transformer models** are popular choices for handling sequential dependencies in data.

# SPP: An example

- **Part-of-Speech (PoS) tagging**



- The **objective** is to learn a model that can predict the most likely sequence of Part-of-Speech tags given an input sequence.

# Conditional Random Fields

- **Conditional Random Fields (CRFs)** are a type of probabilistic graphical model used for **modeling sequential data**, and they have been particularly effective in the context of Named Entity Recognition (NER).
- Undirected (discriminative) graphical models, trained to **maximize conditional probability**  $P(Y|X)$  of output (sequence)  $Y$  given input (sequence)  $X$ .

# CRFs and NER (1)

- **Sequence labeling**

- In NER, the task is often formulated as a **Sequence Prediction Problem**.
- Each word in a sequence is assigned a **label** indicating its entity type (e.g., person, organization, location) or a label indicating that it is not part of any entity (often labeled as "O" for "outside").

- **Feature representation**

- CRFs model the **conditional probability** of a sequence of labels given the input sequence.
- To do this, **features** are defined based on the **observed data**.
- Features can include information about the **current word**, **neighboring words**, their **labels**, and other **contextual information**.

# CRFs and NER (2)

- **Objective function**

- The model is **trained** to maximize the likelihood of the correct sequence labels given the input data.
- The objective function in CRFs aims to **capture dependencies** between labels, ensuring that the predicted labels form coherent and meaningful sequences.

- **Inference**

- During inference (**prediction**), CRFs use algorithms like the **Viterbi algorithm** to find the most likely sequence of labels given the observed data.
- This involves considering both the **individual probabilities of labels** and the **transition probabilities** between them.

# CFRs: Observations and labels (1)

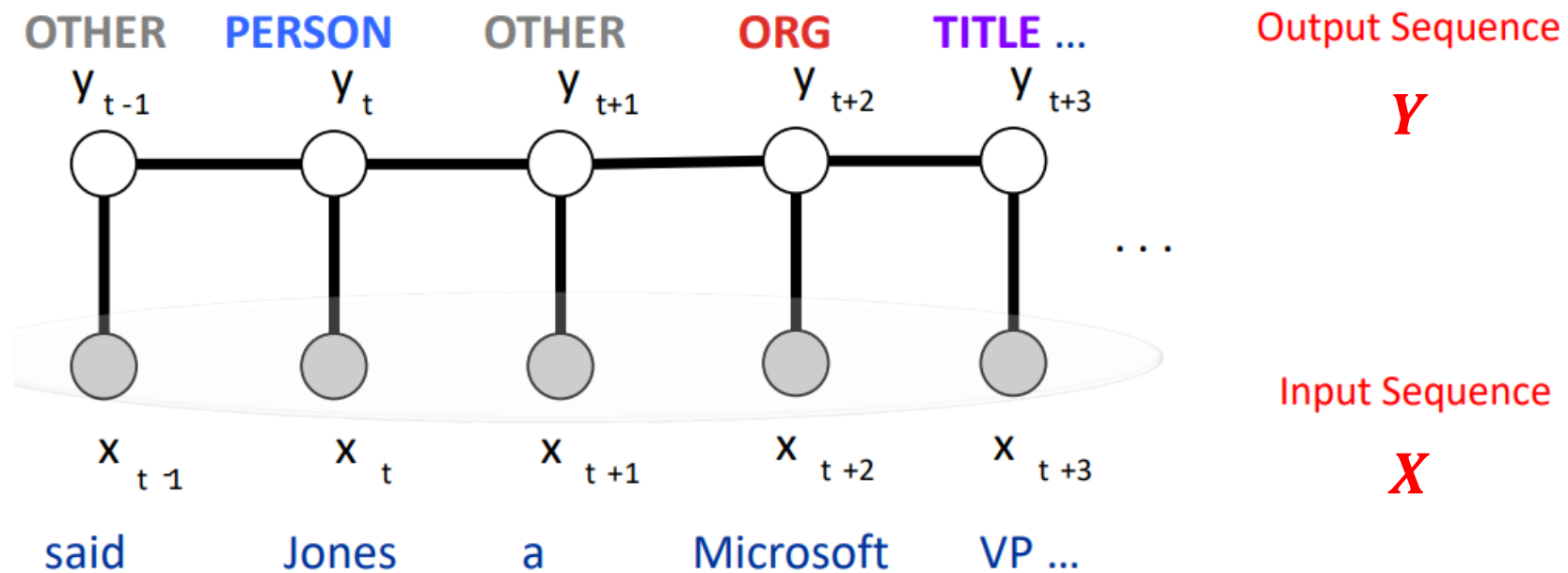
- **Observations**  $X$ :

- The input features, often representing the sequence you want to label.
- $X = \{x_1, x_2, \dots, x_n\}$
- For NER, it could be the **words** in a sentence.

- **Labels**  $Y$ :

- The output labels you want to assign to each element in the sequence.
- $Y = \{y_1, y_2, \dots, y_n\}$
- For NER, it could be **labels** like "person," "organization," etc.

# CFRs: Observations and labels (2)





# CFRs: Conditional probability (1)

- CRFs model the **conditional probability** of a label sequence given the input sequence.

$$P(Y \mid X)$$

- The **basic CRF formula** is defined as follows:

$$P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^n \exp \left( \sum_{j=1}^m \lambda_j \cdot f_j(y_i, y_{i-1}, X, i) \right)$$

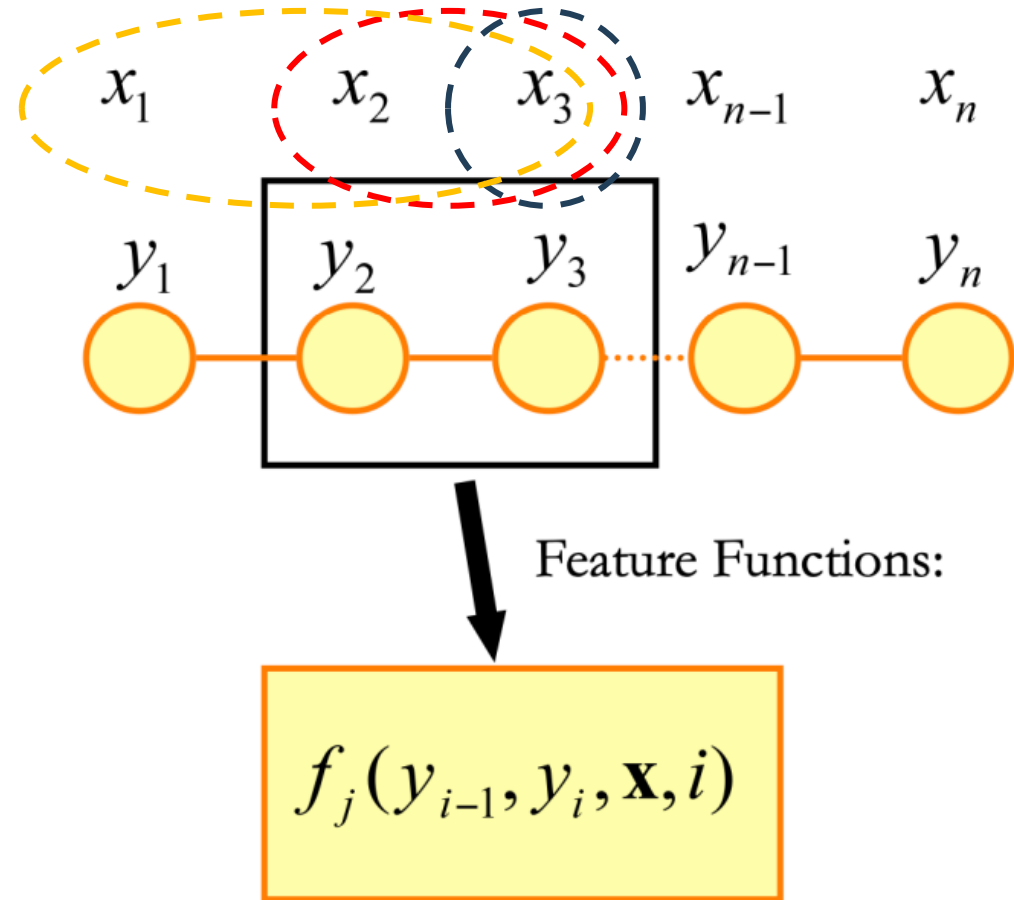
## CFRs: Conditional probability (2)

$$P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^n \exp \left( \sum_{j=1}^m \lambda_j \cdot f_j(y_i, y_{i-1}, X, i) \right)$$

- $Z(X) \rightarrow$  the **normalization term**, ensuring that the probabilities sum up to 1.
- $\lambda_j \rightarrow$  the **model parameters** to be learned.
- $f_j(y_i, y_{i-1}, X, i) \rightarrow$  **feature functions** that **capture dependencies** between labels.

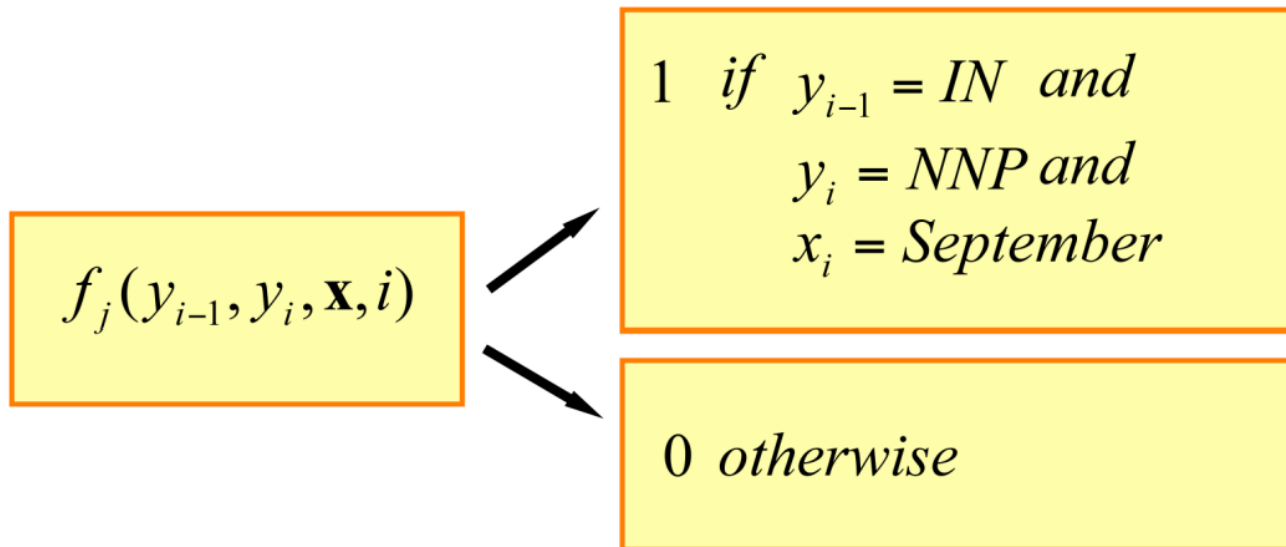
# CFRs: Feature functions (1)

- **Feature functions**  $f_j(y_i, y_{i-1}, X, i)$  are designed based on the considered task and capture relevant information for making labeling decisions.
- They can involve **pairwise interactions** between labels and neighboring observations.
  - $y_i$  and  $y_{i-1}$  are the current and previous labels.
  - $X$  is the input sequence.
  - $i$  is the position in the sequence.



# CFRs: Feature functions (2)

Express some characteristic of the empirical distribution  
that we wish to hold in the model distribution



# A parenthesis: Hidden Markov Models

- Both HMMs and CRFs are types of **probabilistic graphical models** used for Sequence Prediction tasks, but they have different structures and modeling assumptions.
- In an HMM, a **sequence of observations** is assumed to be **generated by an underlying sequence of hidden states**.
  - **Transition probabilities** describe the likelihood of moving from one hidden state to another.
  - **Emission probabilities** describe the likelihood of observing a particular observation given a hidden state.

# CFRs: Feature functions (3)

- **Pairwise potential (transition-like probability)**

- In CRFs, the transition-like probability is captured by the feature functions involving **neighboring labels**.
  - A feature function might look like this:  $f_{\text{transition}}(y_i, y_{i-1})$ .
  - E.g., the probability of transitioning from a location to a person.
- The **weights** ( $\lambda$ ) associated with these features are learned during training.
  - The larger the weight, the more influence that feature has on the model.

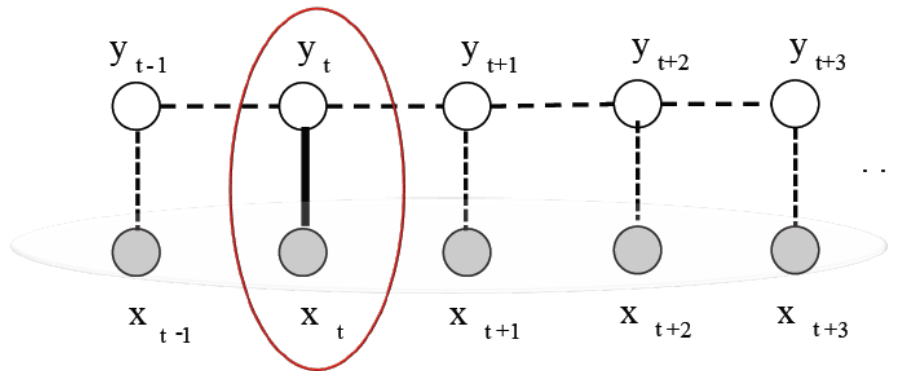
- **Unary potential (emission-like probability)**

- In CRFs, the emission-like probability is captured by feature functions involving the **current label** and the **input sequence**.
  - A feature function might look like this:  $f_{\text{emission}}(y_i, x_i)$ .
  - E.g., the probability of observing the word "New York" given that it is a location.

# CFRs: Feature functions (4)

$y = \text{" Other PERS Other Other PERS... "}$

**Emission-like probability**



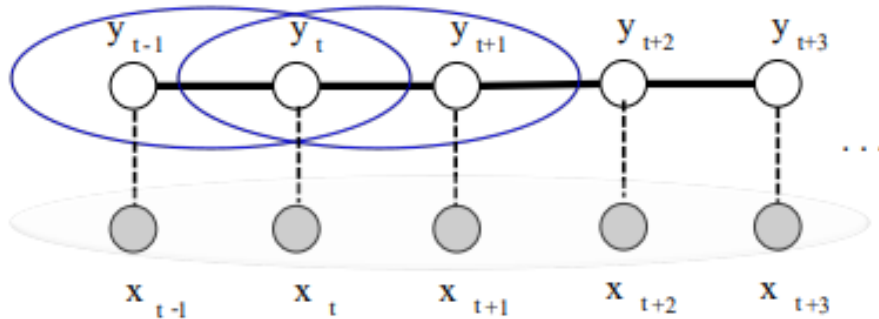
$x = \text{"yesterday John said that Nick was at home"}$

$$f_{\text{emission}}(y_i, x_i) = \begin{cases} 1 & \text{if } y_i = \text{PERS and } x_i = \text{John} \\ 0 & \text{otherwise} \end{cases}$$

# CFRs: Feature functions (5)

$y$  = "Other PERS Other Other PERS Other Other Other "

Transition-like probability



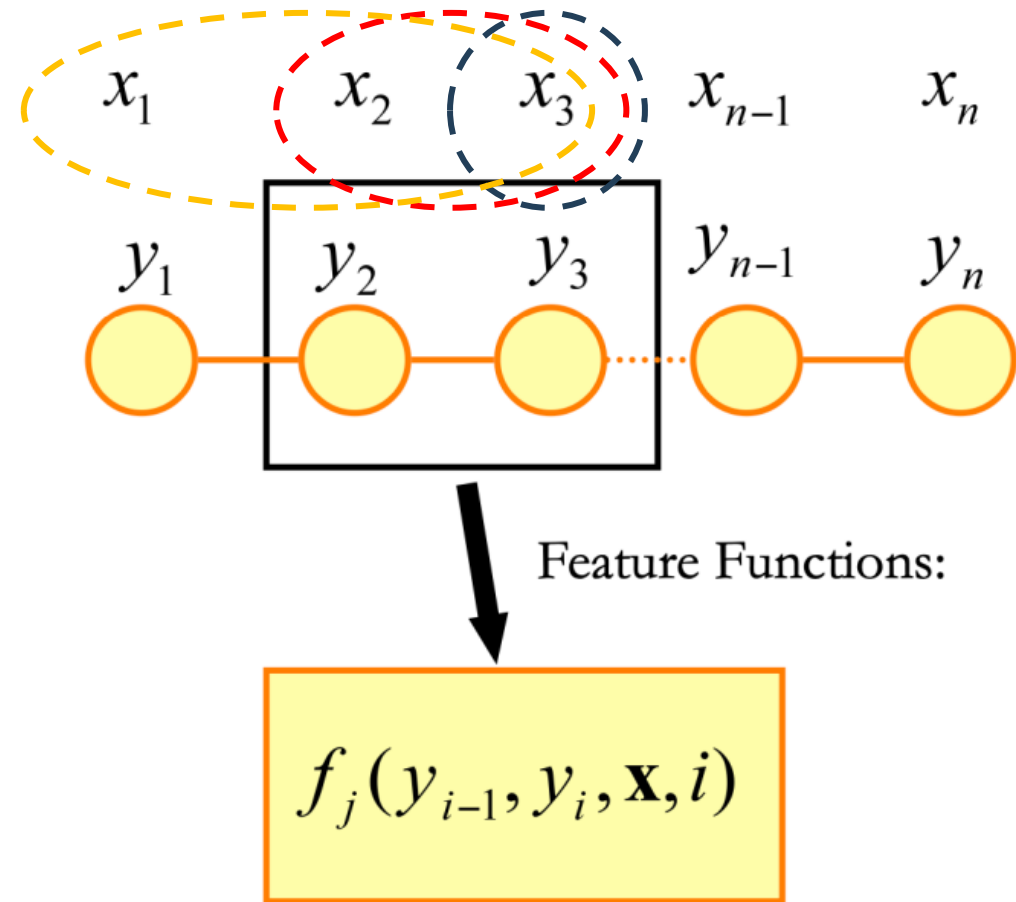
$x$  = "yesterday John said that Nick was at home"

$$f_{\text{transition}}(y_{i-1}, y_i) = \begin{cases} 1 & \text{if } y_{i-1} = \text{OTHER and } x_i = \text{PERS} \\ 0 & \text{otherwise} \end{cases}$$



# CFRs: Feature functions (6)

- Features  $f_j$  capture both transition and emission-like information.
- $\sum_{j=1}^m \lambda_j \cdot f_j$  computes a weighted sum of these features.



# CFRs: Training and inference

- **Training** a CRF involves learning the parameters  $\lambda_j$  from labeled training data.
  - This is typically done using methods like **maximum likelihood estimation** or gradient-based optimization.
  - Once trained, CRFs can be used for **making predictions** on new sequences.
- **Inference**: the **most likely label sequence**  $\hat{Y}$  for a given input sequence  $X$  is found by maximizing the conditional probability:

$$\hat{Y} = \arg \max_Y P(Y|X)$$

- The  $\arg \max$  operation involves selecting the sequence of labels that has the highest probability given the observed input.
  - In the context of NER, this means determining the most likely sequence of named entities for a given input text.

# NER systems

- For updated information about tools and datasets: Jehangir, Basra, Saravanan Radhakrishnan, and Rahul Agarwal. "[A survey on Named Entity Recognition—datasets, tools, and methodologies.](#)" *Natural Language Processing Journal* 3 (2023): 100017

Off-the-Shelf NER Tools Offered by Academia and Industry Projects

NER System	URL
StanfordCoreNLP	<a href="https://stanfordnlp.github.io/CoreNLP/">https://stanfordnlp.github.io/CoreNLP/</a>
OSU Twitter NLP	<a href="https://github.com/aritter/twitter_nlp">https://github.com/aritter/twitter_nlp</a>
Illinois NLP	<a href="http://cogcomp.org/page/software/">http://cogcomp.org/page/software/</a>
NeuroNER	<a href="http://neuroner.com/">http://neuroner.com/</a>
NERsuite	<a href="http://nersuite.nlplab.org/">http://nersuite.nlplab.org/</a>
Polyglot	<a href="https://polyglot.readthedocs.io">https://polyglot.readthedocs.io</a>
Gimli	<a href="http://bioinformatics.ua.pt/gimli">http://bioinformatics.ua.pt/gimli</a>
spaCy	<a href="https://spacy.io/api/entityrecognizer">https://spacy.io/api/entityrecognizer</a>
NLTK	<a href="https://www.nltk.org">https://www.nltk.org</a>
OpenNLP	<a href="https://opennlp.apache.org/">https://opennlp.apache.org/</a>
LingPipe	<a href="http://alias-i.com/lingpipe-3.9.3/">http://alias-i.com/lingpipe-3.9.3/</a>
AllenNLP	<a href="https://demo.allennlp.org/">https://demo.allennlp.org/</a>
IBM Watson	<a href="https://natural-language-understanding-demo.ng.bluemix.net/">https://natural-language-understanding-demo.ng.bluemix.net/</a>
FG-NER	<a href="https://fgner.alt.ai/extractor/">https://fgner.alt.ai/extractor/</a>
Intellexer	<a href="http://demo.intellexer.com/">http://demo.intellexer.com/</a>
Repustate	<a href="https://repustate.com/named-entity-recognition-api-demo/">https://repustate.com/named-entity-recognition-api-demo/</a>
AYLIEN	<a href="https://developer.aylien.com/text-api-demo">https://developer.aylien.com/text-api-demo</a>
Dandelion API	<a href="https://dandelion.eu/semantic-text/entity-extraction-demo/">https://dandelion.eu/semantic-text/entity-extraction-demo/</a>
displaCy	<a href="https://explosion.ai/demos/displacy-ent">https://explosion.ai/demos/displacy-ent</a>
ParallelDots	<a href="https://www.paralleldots.com/named-entity-recognition">https://www.paralleldots.com/named-entity-recognition</a>
TextRazor	<a href="https://www.textrazor.com/named_entity_recognition">https://www.textrazor.com/named_entity_recognition</a>

# NER datasets

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	<a href="https://catalog ldc.upenn.edu/LDC2003T13">https://catalog ldc.upenn.edu/LDC2003T13</a>
MUC-6 Plus	1995	Additional news to MUC-6	7	<a href="https://catalog ldc.upenn.edu/LDC96T10">https://catalog ldc.upenn.edu/LDC96T10</a>
MUC-7	1997	New York Times news	7	<a href="https://catalog ldc.upenn.edu/LDC2001T02">https://catalog ldc.upenn.edu/LDC2001T02</a>
CoNLL03	2003	Reuters news	4	<a href="https://www.clips.uantwerpen.be/conll2003/ner/">https://www.clips.uantwerpen.be/conll2003/ner/</a>
ACE	2000 - 2008	Transcripts, news	7	<a href="https://www ldc.upenn.edu/collaborations/past-projects/ace">https://www ldc.upenn.edu/collaborations/past-projects/ace</a>
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	<a href="https://catalog ldc.upenn.edu/LDC2013T19">https://catalog ldc.upenn.edu/LDC2013T19</a>
W-NUT	2015 - 2018	User-generated text	6/10	<a href="http://noisy-text.github.io">http://noisy-text.github.io</a>
BBN	2005	Wall Street Journal	64	<a href="https://catalog ldc.upenn.edu/LDC2005T33">https://catalog ldc.upenn.edu/LDC2005T33</a>
WikiGold	2009	Wikipedia	4	<a href="https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500">https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500</a>
WiNER	2012	Wikipedia	4	<a href="http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner">http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner</a>
WikiFiger	2012	Wikipedia	112	<a href="https://github.com/xiaoling/figer">https://github.com/xiaoling/figer</a>
HYENA	2012	Wikipedia	505	<a href="https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena/">https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena/</a>
N <sup>3</sup>	2014	News	3	<a href="http://aksw.org/Projects/N3NERNEDNIF.html">http://aksw.org/Projects/N3NERNEDNIF.html</a>
Gillick	2016	Magazine, news, web, etc.	89	<a href="https://arxiv.org/e-print/1412.1820v2">https://arxiv.org/e-print/1412.1820v2</a>
FG-NER	2018	Various	200	<a href="https://fgner.alt.ai/">https://fgner.alt.ai/</a>
NNE	2019	News wire	114	<a href="https://github.com/nickyringland/nested_named_entities">https://github.com/nickyringland/nested_named_entities</a>
GENIA	2004	Biology and clinical text	36	<a href="http://www.geniaproject.org/home">http://www.geniaproject.org/home</a>
GENETAG	2005	MEDLINE	2	<a href="https://sourceforge.net/projects/bioc/files/">https://sourceforge.net/projects/bioc/files/</a>
FSU-PRGE	2010	PubMed and MEDLINE	5	<a href="https://julielab.de/Resources/FSU_PRGE.html">https://julielab.de/Resources/FSU_PRGE.html</a>
NCBI-Disease	2014	PubMed	1	<a href="https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/">https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/</a>
BC5CDR	2015	PubMed	3	<a href="http://bioc.sourceforge.net/">http://bioc.sourceforge.net/</a>
DFKI	2018	Business news and social media	7	<a href="https://dfki-lt-re-group.bitbucket.io/product-corpus/">https://dfki-lt-re-group.bitbucket.io/product-corpus/</a>

# NER: Recent trends

- Conditional Random Fields have been widely used in NER tasks due to their ability to model **complex dependencies in sequential data**.
- However, more recently, **deep learning approaches**, such as **Recurrent Neural Networks** (RNNs) and **Transformer models**, have also gained popularity in NER tasks, often outperforming traditional CRF-based models.
- The choice between CRFs and deep learning methods may depend on the specific requirements of the NER task and the available data.

# Evaluating NER (1)

- **Exact match**

New Frozen Boutique to Open at Disney's Hollywood Studios.  
Can't wait 😞

- **Partial match**

New Frozen Boutique to Open at Disney's Hollywood Studios.  
Can't wait 😞

## Evaluating NER (2)

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}.$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

# Evaluating NER (3)

- **Use of domain-specific metrics**

- Depending on the application, one might need to define and use **domain-specific metrics**.
- For example, in biomedical NER, one might focus on the identification of specific entities like **genes** or **diseases**.

- **Evaluating NER without a ground truth** can be challenging

- **Rule-based evaluation**: a rule-based system can act as a pseudo ground truth for the NER model.
- **Crowdsourcing**: use of crowdsourcing platforms to generate annotations.
- **Semi-supervised learning**.

- **Active learning**: select the most informative samples for manual annotation.

- This involves iteratively training the model on the existing data and querying the user to label instances that the model is uncertain about.

- **Simulation**: simulate the ground truth by introducing synthetic entities into the text data.

- The synthetic data should be representative of the real-world use case.

- **Expert evaluation**: expert opinions can help assess the reasonableness of the model's performance.

- **Evaluation on similar tasks**: use of a labeled dataset for a task closely related to NER.



# Named Entity Linking

---

# Initial hypothesis after NER

- Once we have identified entity mentions, we must (and we have an initial **hypothesis** of their types), we must “disambiguate” them.

location

New Frozen Boutique to Open at Disney's Hollywood Studios.

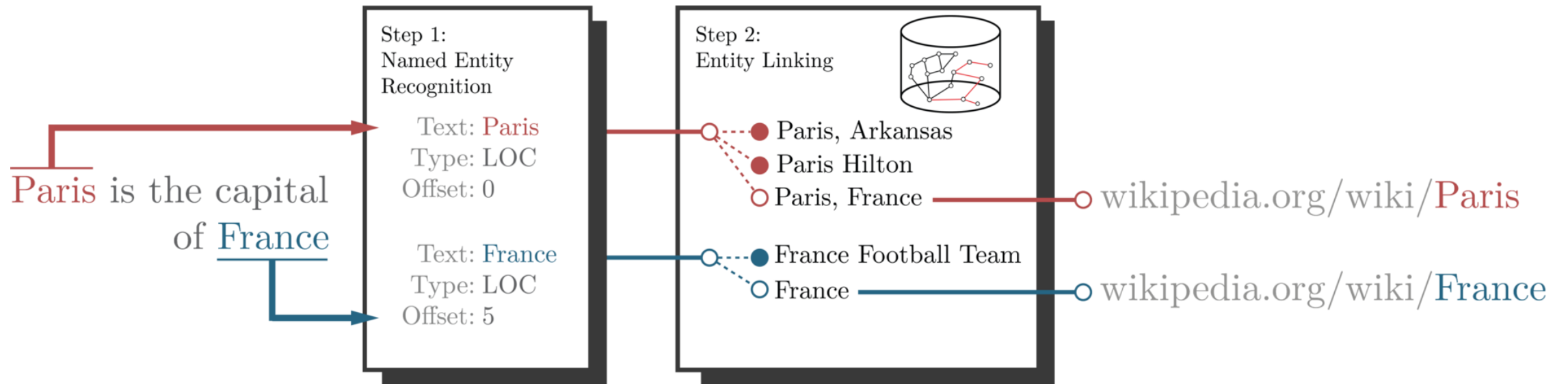
product

organization

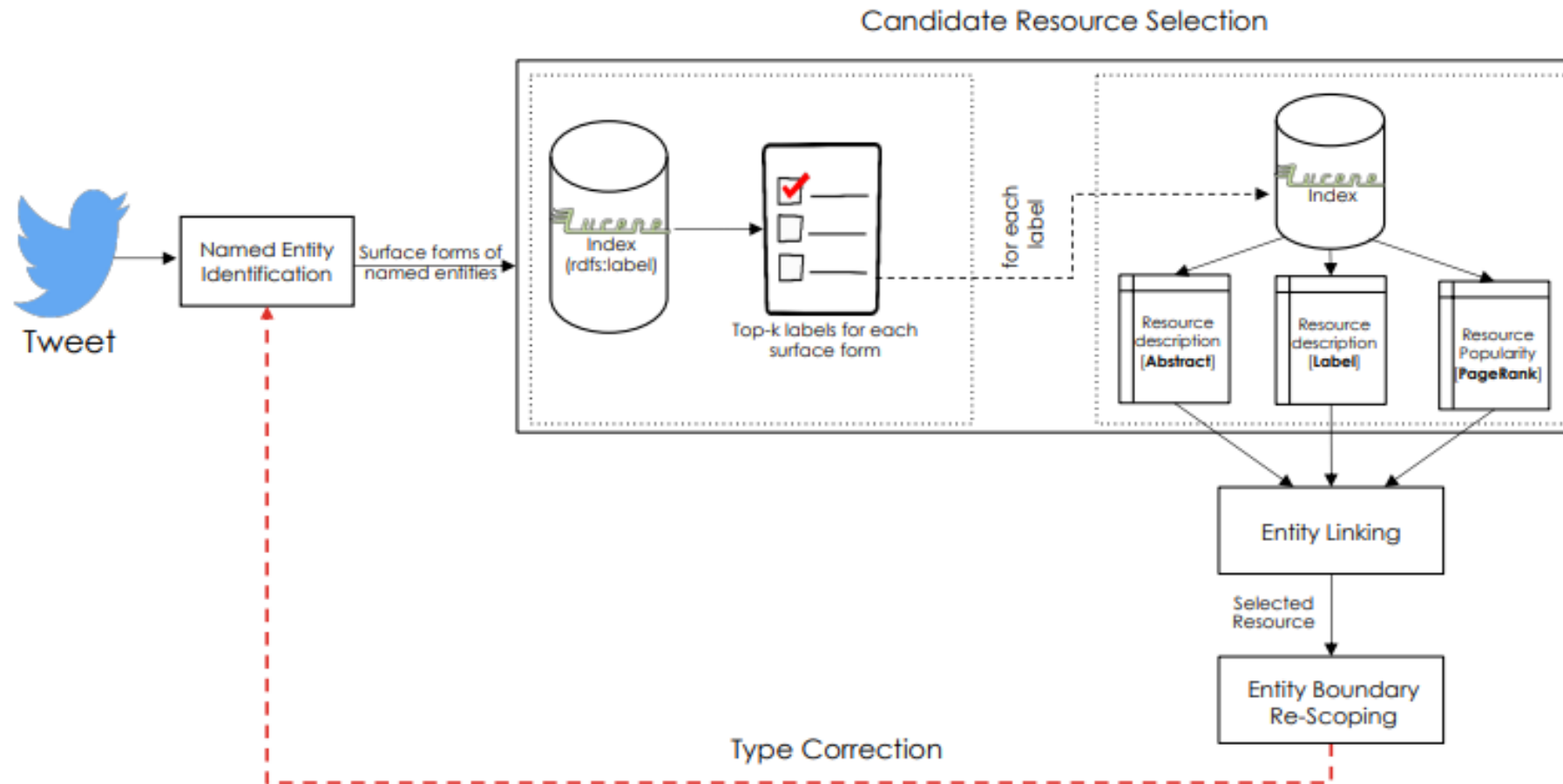
# Deepening NEL

- **Named Entity Linking (NEL)** is the task of associating mentions of Named Entities in text with **unique identifiers** or **entries** in a **knowledge base** (or in a reference database).
- It involves **determining the correct entity that a mention refers to** (*this is overlapping with Named Entity Disambiguation*), considering potential candidates from a knowledge base.

# NEL: Example



# IR-based NEL



# Entity resources

- **Wikidata**

- A free and open knowledge base that can be used for linking entities mentioned in text to their corresponding entries in Wikidata.
  - <https://www.wikidata.org/wiki/?uselang=en>

- **DBpedia**

- Extracts structured information from Wikipedia and provides **RDF triples**. It can be used for linking entities mentioned in text to their DBpedia entries.
  - <https://www.dbpedia.org/>

- **YAGO**

- A large knowledge base with **entities** and **relations**, which can be used for Named Entity Linking tasks.
  - <https://yago-knowledge.org/>

# NEL identifiers

- In Named Entity Linking (NEL), **identifiers** play a crucial role in linking entity mentions in text to specific entries in a knowledge base or reference database.
- Identifiers **uniquely identify** entities and allow systems to establish a connection between the textual mention and the corresponding entity in a structured knowledge representation.
- Commonly used types of identifiers in NEL include **URLs**, **RDF triples**, and **other unique identifiers**.

# NEL identifiers: Characteristics

- **Uniqueness**: Identifiers should be unique within the context of the knowledge base or reference database to ensure that each entity has a distinct identifier.
- **Stability**: Ideally, identifiers should remain stable over time to ensure the persistence of links between entity mentions and knowledge base entries.
- **Accessibility**: Identifiers should be accessible and retrievable, allowing systems to fetch additional information about the linked entities.



# NEL identifiers: URLs and RDF triplets

## URLs (Uniform Resource Locators)

- **URLs** are web addresses that uniquely identify resources on the internet.
- In the context of NEL, entities are often associated with **URLs that point to their corresponding pages** or entries in online knowledge bases.
- For example:
  - Entity mention: Barack Obama
  - URL identifier:  
[https://en.wikipedia.org/wiki/Barack\\_Obama](https://en.wikipedia.org/wiki/Barack_Obama)

## RDF (Resource Description Framework) Triplets

- **RDF** is a standard for representing information about **resources** on the Web in the form of **triples**.
- *Subject-predicate-object* statements.

# *A parenthesis: RDF (1)*

- **RDF** represents information as triples, where each triple comprises three components: subject, predicate, and object.
  - **Subject (S)**: The resource being described.
  - **Predicate (P)**: The property or attribute of the resource.
  - **Object (O)**: The value or another resource related to the subject.
- **Example.** Let's consider the statement: "The sky is blue."
  - **Subject (S)**: The sky
  - **Predicate (P)**: hasColor
  - **Object (O)**: blue

# *A pharentesis*: RDF (2)

- **RDF graph**

- As more triples are added, they form a graph where nodes represent resources, and edges represent relationships (predicates) between them.

- **Real-world application**

- RDF is widely used for linking and integrating data on the web.
- It enables the creation of a **Linked Data ecosystem** where information from different sources is connected through RDF triples.
  - Mountantonakis, Michalis, and Yannis Tzitzikas. "**Large-scale semantic integration of linked data: A survey.**" *ACM Computing Surveys* (CSUR) 52.5 (2019): 1-40.

# *A parenthesis: RDF (3)*



<https://neo4j.com/blog/neo4j-rdf-graph-database-reasoning-engine/>

# NEL identifiers: RDF triplets

- When **RDF** is used for Named Entity Linking:
  - The **subject** is the **entity identifier**.
  - The **predicate** is a specific **property**.
  - The **object** is the **value** connected to the property.
- **Example:**
  - Entity identifier (Subject): [http://dbpedia.org/resource/Barack\\_Obama](http://dbpedia.org/resource/Barack_Obama)
  - Property (Predicate): <http://xmlns.com/foaf/0.1/name>
  - Value (Object): "Barack Obama"

# NEL identifiers: Wikidata and DBpedia IDs

## Wikidata identifiers

- Wikidata, a free and open knowledge base, uses its own unique identifiers for entities.
- These identifiers are often used in NEL tasks that involve linking to Wikidata.
- For example:
  - Entity Identifier: Q76

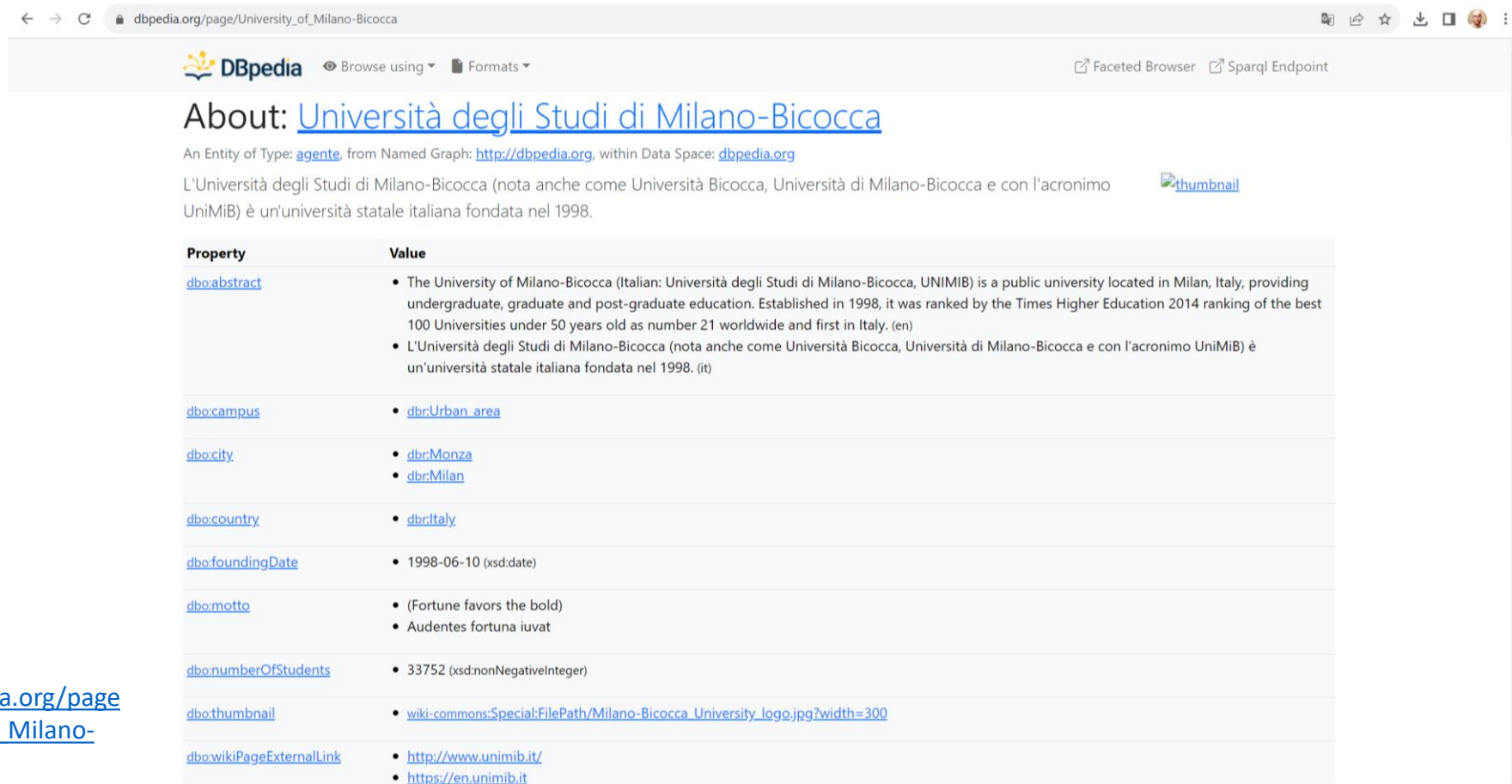
## DBpedia identifiers

- DBpedia, which extracts structured content from Wikipedia, uses its own identifiers for entities.
- These identifiers are commonly used in NEL tasks that involve linking to DBpedia.
- For example:
  - Entity identifier:  
[http://dbpedia.org/resource/Barack\\_Obama](http://dbpedia.org/resource/Barack_Obama)

# DBpedia

- **DBpedia** is a community-driven project that extracts structured information from Wikipedia and makes it available in a machine-readable format.
  - <https://www.dbpedia.org/>
- The data in DBpedia is organized using the **Resource Description Framework (RDF)**.

# DBpedia: An example




← → ↻ dbpedia.org/page/University\_of\_Milano-Bicocca

DBpedia Browse using Formats Faceted Browser Sparql Endpoint

## About: [Università degli Studi di Milano-Bicocca](#)

An Entity of Type: [agente](#), from Named Graph: <http://dbpedia.org>, within Data Space: [dbpedia.org](#)

L'Università degli Studi di Milano-Bicocca (nota anche come Università Bicocca, Università di Milano-Bicocca e con l'acronimo UniMiB) è un'università statale italiana fondata nel 1998. 

Property	Value
<a href="#">dbo:abstract</a>	<ul style="list-style-type: none"><li>The University of Milano-Bicocca (Italian: Università degli Studi di Milano-Bicocca, UNIMiB) is a public university located in Milan, Italy, providing undergraduate, graduate and post-graduate education. Established in 1998, it was ranked by the Times Higher Education 2014 ranking of the best 100 Universities under 50 years old as number 21 worldwide and first in Italy. (en)</li><li>L'Università degli Studi di Milano-Bicocca (nota anche come Università Bicocca, Università di Milano-Bicocca e con l'acronimo UniMiB) è un'università statale italiana fondata nel 1998. (it)</li></ul>
<a href="#">dbo:campus</a>	<ul style="list-style-type: none"><li><a href="#">dbr:Urban_area</a></li></ul>
<a href="#">dbo:city</a>	<ul style="list-style-type: none"><li><a href="#">dbr:Monza</a></li><li><a href="#">dbr:Milan</a></li></ul>
<a href="#">dbo:country</a>	<ul style="list-style-type: none"><li><a href="#">dbr:Italy</a></li></ul>
<a href="#">dbo:foundingDate</a>	<ul style="list-style-type: none"><li>1998-06-10 (xsd:date)</li></ul>
<a href="#">dbo:motto</a>	<ul style="list-style-type: none"><li>(Fortune favors the bold)</li><li>Audentes fortuna iuvat</li></ul>
<a href="#">dbo:numberOfStudents</a>	<ul style="list-style-type: none"><li>33752 (xsd:nonNegativeInteger)</li></ul>
<a href="#">dbo:thumbnail</a>	<ul style="list-style-type: none"><li><a href="#">wiki-commons:Special:FilePath/Milano-Bicocca_University_logo.jpg?width=300</a></li></ul>
<a href="#">dbo:wikiPageExternalLink</a>	<ul style="list-style-type: none"><li><a href="http://www.unimib.it/">http://www.unimib.it/</a></li><li><a href="https://en.unimib.it">https://en.unimib.it</a></li></ul>

[https://dbpedia.org/page/University\\_of\\_Milano-Bicocca](https://dbpedia.org/page/University_of_Milano-Bicocca)



# NEL: Techniques

- Various techniques and approaches are used for Named Entity Linking, and they can be broadly categorized into **Text-based NEL** and **Graph-based NEL**.
- **Text-based Named Entity Linking**
  - It relies on the text representation of the entity mentions and does not explicitly use graph-based representations of entities.
- **Graph-based Named Entity Linking**
  - It explicitly use graph-based representations of entities.

# Text-based NEL

- The seminal work by **Cucerzan in 2007** proposed one of the first entity linking systems that appeared in the literature, and tackled the task of **wikification**, linking textual mentions to **Wikipedia pages**.
  - Silviu Cucerzan. 2007. **Large-Scale Named Entity Disambiguation Based on Wikipedia Data**. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

# Text-based NEL: Techniques

- **String matching**

- Simple string matching techniques involve comparing the mention in the text to entries in a knowledge base.
- Exact matches or approximate matches (e.g., Levenshtein distance) can be used to link entities.

- **Entity co-occurrence**

- Entities that often co-occur in a similar context may be linked together.
- This method leverages statistical patterns in the data to establish connections between entities.

- **Machine Learning models**

- Supervised learning models, such as Support Vector Machines (SVMs), Conditional Random Fields (CRFs), and more recently, deep learning models, can be trained on labeled data to predict entity links based on contextual information.

- **Embedding models**

- Embedding techniques, such as Word Embeddings (e.g., Word2Vec, GloVe) and Contextual Embeddings (e.g., BERT, ELMO), can capture semantic relationships between words and entities, aiding in linking.

# Text-based NEL: String matching (1)

- *KB* (DBpedia) articles (`rdfs:labels`)
- Given an entity  $e_j$ , we compute a **KB score** w.r.t. each resource  $c_k$

$$KB(e_j, c_k) = \alpha \cdot \underbrace{lex(e_j, l_{c_k})}_{\text{lexical similarity}} + (1 - \alpha) \cdot \underbrace{cov(e_j, c_k)}_{\text{coverage}}$$

**lexical similarity** between an **entity**  $e_j$  and the **label** of a candidate resource label  $l_{c_k}$

**coverage** based on the **coherence** of an entity  $e_j$  w.r.t KB abstracts and the **popularity** of a candidate resource  $c_k$

# Text-based NEL: String matching (2)

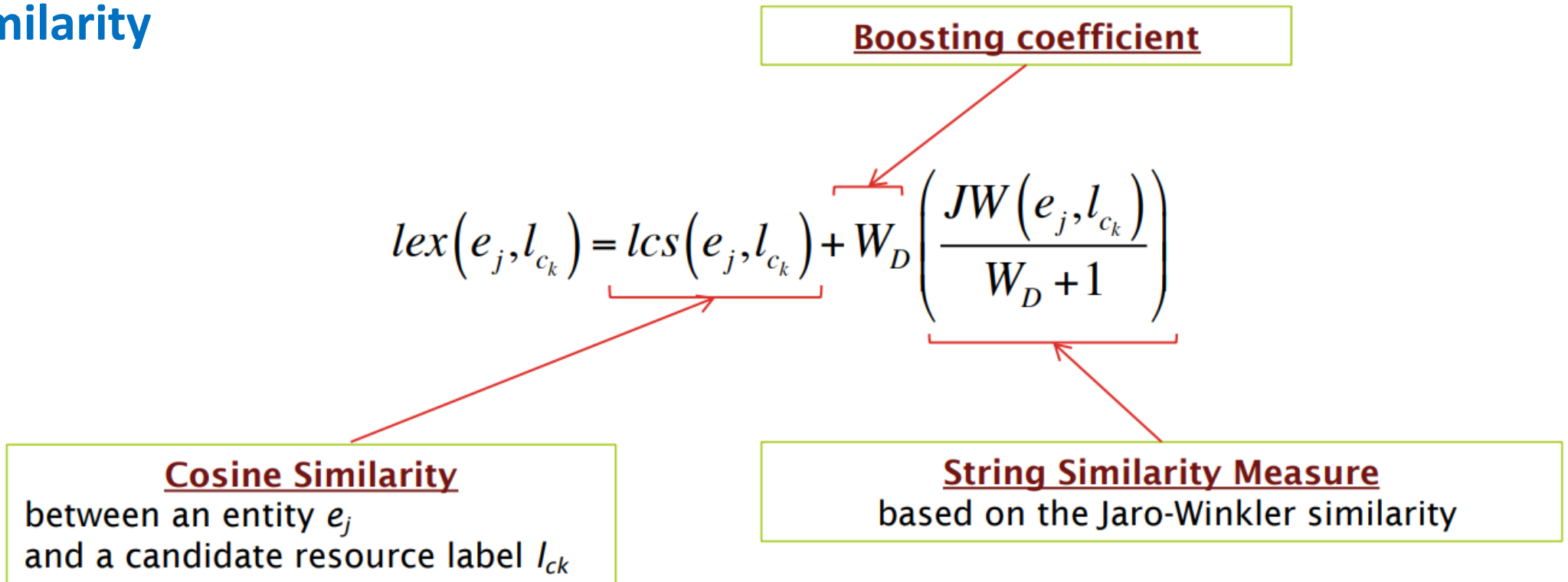
- **Similarity**

$$lex(e_j, l_{c_k}) = \underbrace{lcs(e_j, l_{c_k})}_{\text{Cosine Similarity}} + \underbrace{W_D \left( \frac{JW(e_j, l_{c_k})}{W_D + 1} \right)}_{\text{String Similarity Measure based on the Jaro-Winkler similarity}}$$

**Boosting coefficient**

**Cosine Similarity**  
between an entity  $e_j$   
and a candidate resource label  $l_{c_k}$

**String Similarity Measure**  
based on the Jaro-Winkler similarity



# Text-based NEL: String matching (3)

- Coverage

$$\text{cov}(e_j, l_{c_k}) = \underbrace{\cos(e_j^*, a_{c_k})}_{\text{Cosine Similarity}} + \underbrace{R(c_k)}_{\text{Page Rank}}$$

**Cosine Similarity**  
between an entity context  $e_j^*$   
and a resource abstract  $a_{c_k}$

**Page Rank**  
based on popularity in the KB

# Text-based NEL: String matching (4)

- **Ranking resources** according to  $KB(e_j, c_k)$ .
- Selecting the optimal resource  $c_k$  for each entity  $e_j$ .

$$c^* = \arg \max[KB(e_j, c_k)], \forall e_j$$

- The final type is selected according to the `rdf:type` of the optimal resource  $c^*$ .

# Text-based NEL: String matching (5)

- **Entity boundary re-scoping**
  - Post-processing on the identified entities for reducing noise

I bought the StarWars t-shirt



I bought the StarWars t-shirt



# Graph-based NEL

- **Graph-based Named Entity Linking (NEL)** is an approach to entity linking that leverages the **structure** and **relationships** present in a **knowledge graph**.
- A **knowledge graph**, also known as a semantic network, represents a network of real-world entities – i.e., objects, events, situations, or concepts – and illustrates the relationship between them.
  - This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge “graph.”

# Graph-based NEL: Key components (1)

- **Knowledge graph representation**

- A **knowledge graph** is used to represent entities and their relationships.
- In these graphs, entities are nodes, and relationships between entities are represented as edges.

- **Mention detection**

- Similar to text-based NEL, the process starts with **identifying** and **extracting** Named Entity mentions from the input text using Named Entity Recognition (**NER**).

- **Candidate generation**

- For each identified mention, a **set of candidate entities** is generated from the knowledge graph.
- These candidates are **potential matches** for the given mention.

- **Graph algorithms**

- Graph-based algorithms are applied to analyze the **connectivity** and **relationships** between entities in the knowledge graph.

# Graph-based NEL: Key components (2)

- **Integration with textual features**

- Some graph-based NEL systems may also incorporate **textual features** (e.g., context embeddings or entity embeddings), to enhance the disambiguation process.

- ***Disambiguation***

- The model assigns the **most likely entity link** for each mention based on the graph-based scores and relationships.
- The disambiguation process takes into account the **global context and relationships** between entities in the graph.

- **Evaluation**

- The performance of the graph-based NEL system is evaluated based on how accurately it links entity mentions to their correct entities in the knowledge graph.
- Evaluation metrics include precision, recall, and F1-score.