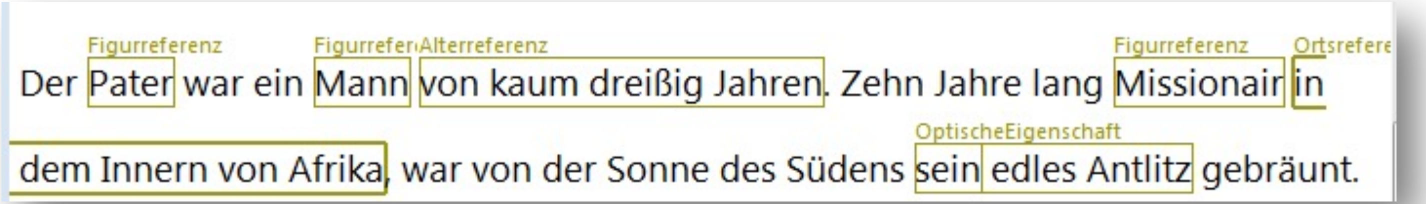
Information Extraction

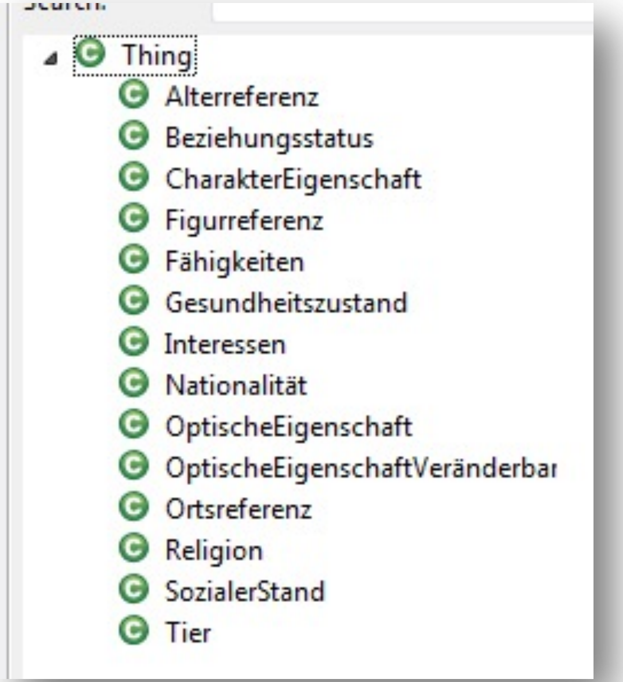
Named Entity Recognition

What is Named Entity Recognition

- Given the text with some syntactical processing
- And an ontology with pre-defined classes of interesting entities

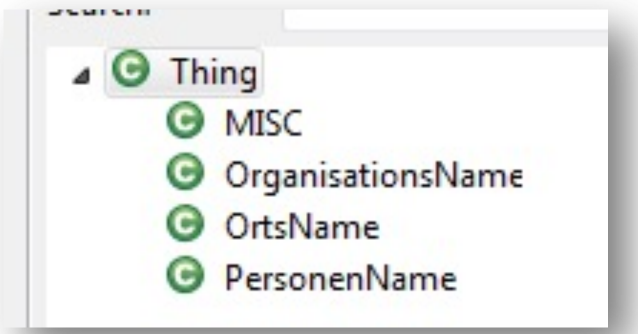
- Example:





What is Named Entity Recognition

- In literature you will usually find "Named" Entity Recognition using 4 classes:
  - Persons ("Bill Gates")
  - Geo-political-locations (GPE) ("Berlin")
  - Organisations ("Google")
  - Misc (e.g. title of a book)



- Depending on domain, there might be more:
  - Measurements
  - Temporal Expressions
  - Drug names
  - Diseases
  - …

What is Named Entity Recognition

- As you might already expect, there are yet again successful methods based on rules and based on machine learning

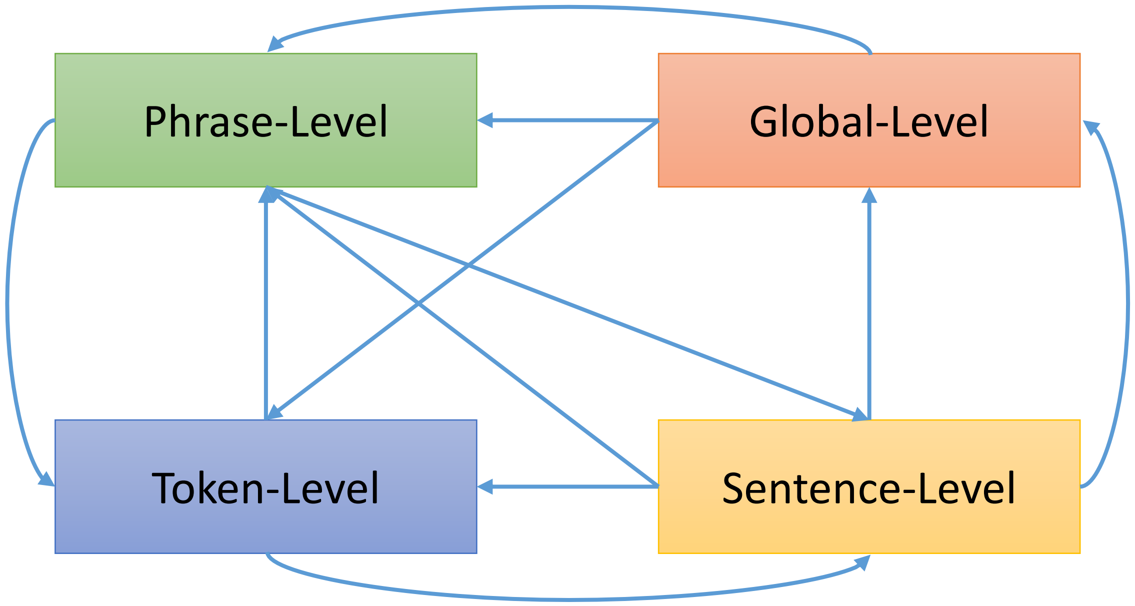
- the next slides will present both approaches

Information Extraction

Rule-Based Approaches

Rule-Based Approaches

- Rule-based approaches have the advantage, that they can operate on **different representations** of the text, in arbitrary order



Rule-Based Approaches

- The token Level: Token-Level

- Can make use of (domain-specific) lists (called gazetteers)

A	B	C	D	E
Liste 1			Liste 2	
Name	x		Name	x
Dumbo	x		Dumbo	x
Susi	x		Dagobert	x
Strolch	x		Minnie	x
Dagobert	x		Micky	x
Minnie	x		Tick	x
Micky	x		Trick	x
			Track	x

- Can use taxonomies (e.g. WordNet or GermaNet)

- Can use common Pre/Suffixes for disambiguation:
  - E.g. in German the suffix „…keit“ usually does not correspond to a name

Rule-Based Approaches

- The phrase Level: Phrase-Level

- Can make use of (domain-specific) templates, in which names occur

  - Prof. Dr. <Name>
  - … we are arriving in/at <Location>
  - <Name> has founded <Organisation>
  - …

Rule-Based Approaches

- The sentence Level:

Sentence-Level

- Can make use of (domain-specific) frames, in which names occur
  - It is usually persons, who are speaking
  - Usually persons who are born

Subject of some Verbs

  - Usually persons who are given a present

Object of some Verbs

Rule-Based Approaches

- The global Level: Global-Level

- If we have rules on any previous level with very high confidence, we can
  - Use the results of previous rules, to „detect“ that name (and potential declinations, e.g. Markus‘) everywhere in a given context

Rule-Based Approaches

- A good rule-based algorithm does now face the task to:
  - Model all individual stages
  - Find an order (might contain loops) in which the rules are applied
  - Potentially clean the result from the previous stage

Rule-Based Approaches

- Taken from the GATE Framework, ANNIE (“a nearly new IE system ”)
- A bunch (>1000) of rules, that are formulated in the language JAPE:

- Person
  - gender: male, female
- Location
  - locType: region, airport, city, country, county, province, other
- Organization
  - orgType: company, department, government, newspaper, team, other
- Money
- Percent
- Date
  - kind: date, time, dateTime
- Address
  - kind: email, url, phone, postcode, complete, ip, other
- Identifier
- Unknown

Rule-Based Approaches

- Example in JAPE:

```
Rule: PersonFirstContext
Priority: 30
// Anne and Kenton

(FIRSTNAME):person1
(
 {Token.string == "and"}
)
({Token.orth == upperInitial, Token.length != "1"})
:person2
 -->
{ ...
```

- Matches on text spans where one person was found already, and is followed by the token “and” and the next token is in upper case and longer than 1

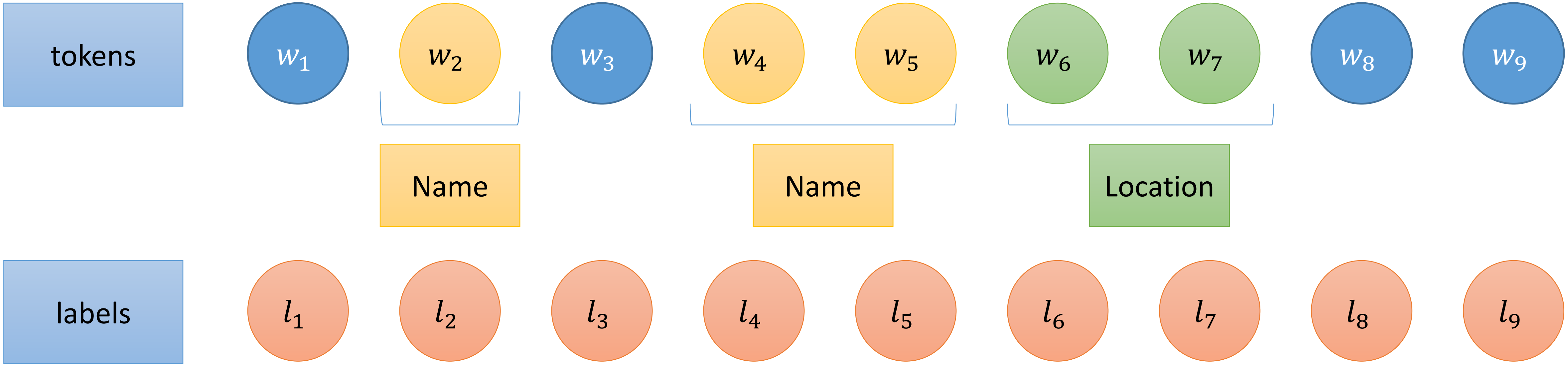
➔ ca. 1200 lines of code for 4 labels

Information Extraction

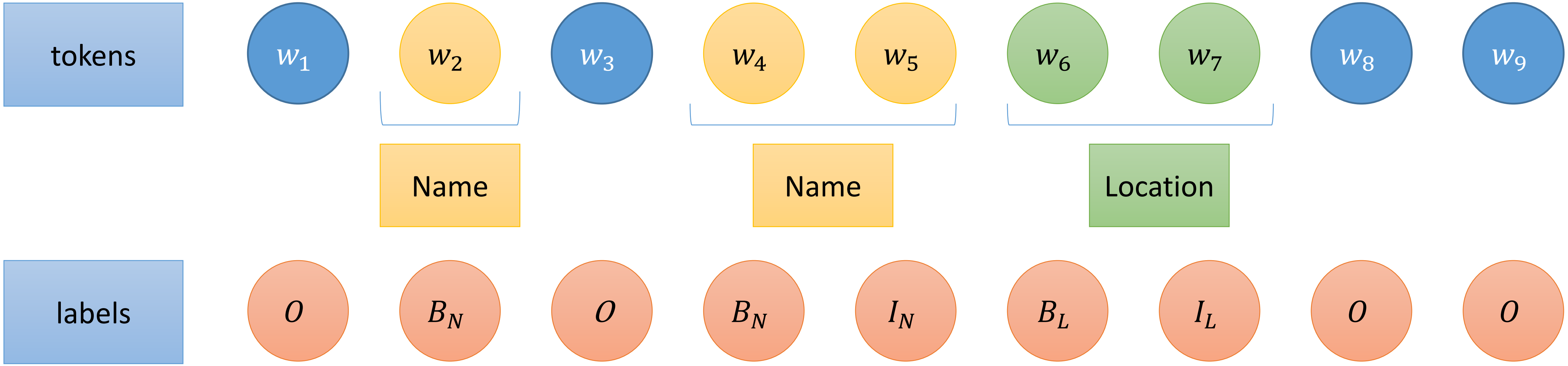
Machine Learning Based Approaches

(Named) Entity Recognition using Machine Learning

- Recall, that we are trying to find some **spans of text**
- Using a very easy trick, we can cast this task into a conventional sequence classification task



(Named) Entity Recognition using Machine Learning



➔ Instead of predicting the entire spans at once, we model a span using its begin (**B**) and its inner body (**I**). All tokens that are not of interest to us will receive the label **O** („out of span“)

➔ Usually called BIO-encoding

(Named) Entity Recognition using Machine Learning

- So recall, that we can make use of the BIO-encoding in order to convert the task of span detection into a conventional sequence classification task.

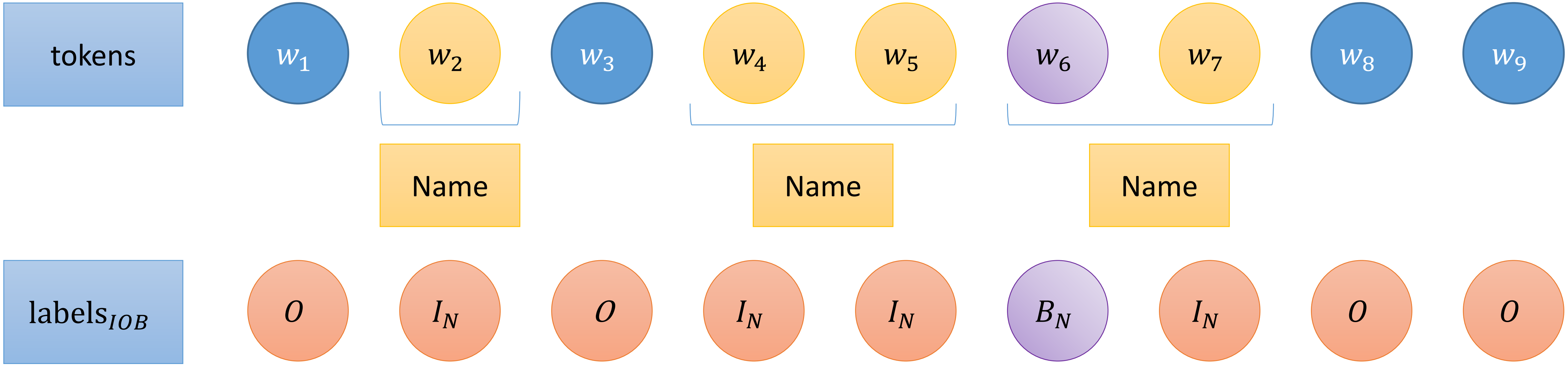
- BIO, will create 2 labels for every class $(B_{\text{class}}, I_{\text{class}})$ as well as one additional label O for tokens that are not of interest

➔ $2k + 1$ labels for k different classes

(Named) Entity Recognition using Machine Learning

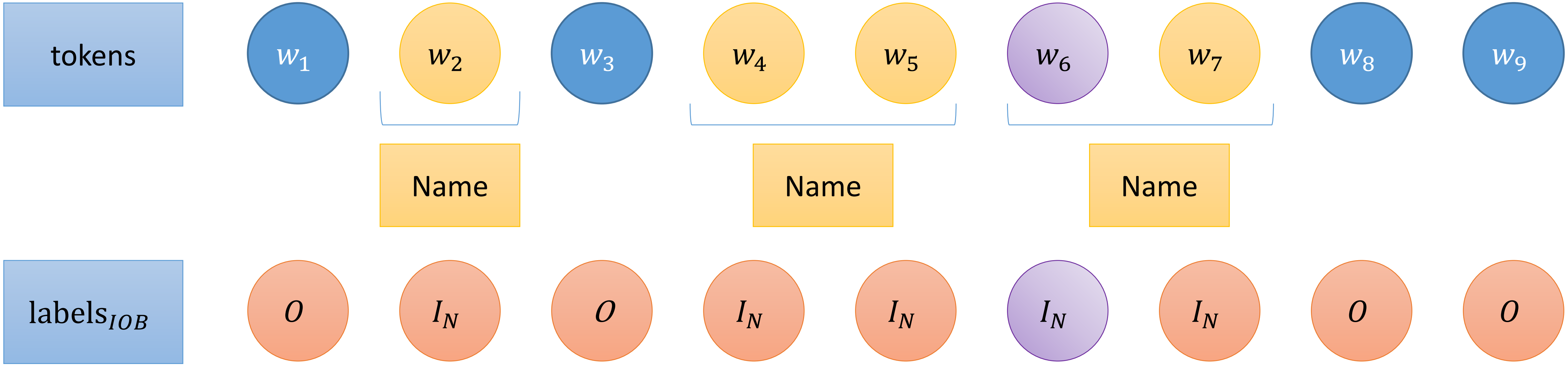
- Other encoding schemas:

- IOB: Only use the „B“, if absolutely necessary!!



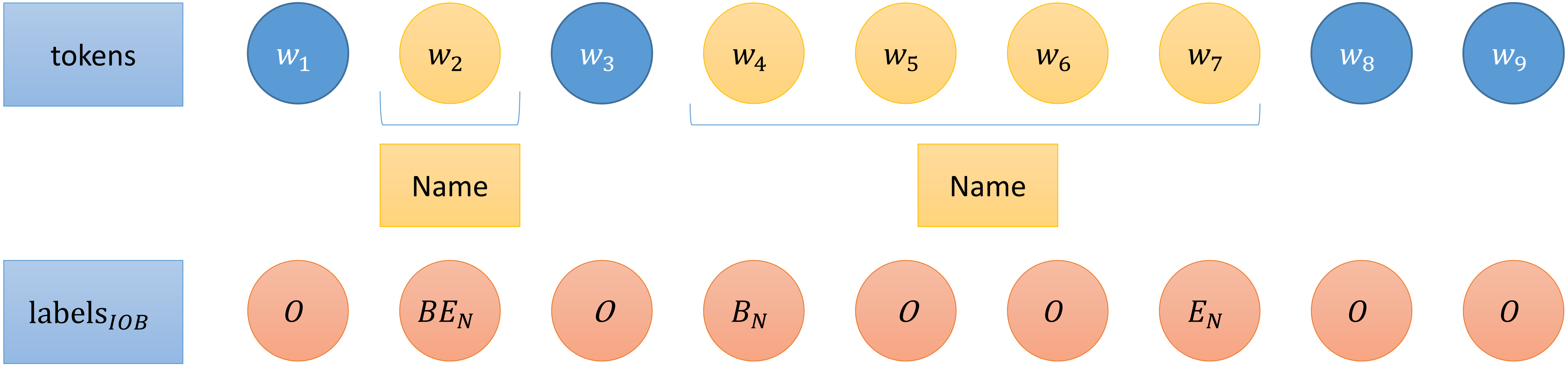
(Named) Entity Recognition using Machine Learning

- Other encoding schemas: IO – do not use B at all!
  - Lossy (see below, span gets lost)
  - But only $k + 1$ classes!



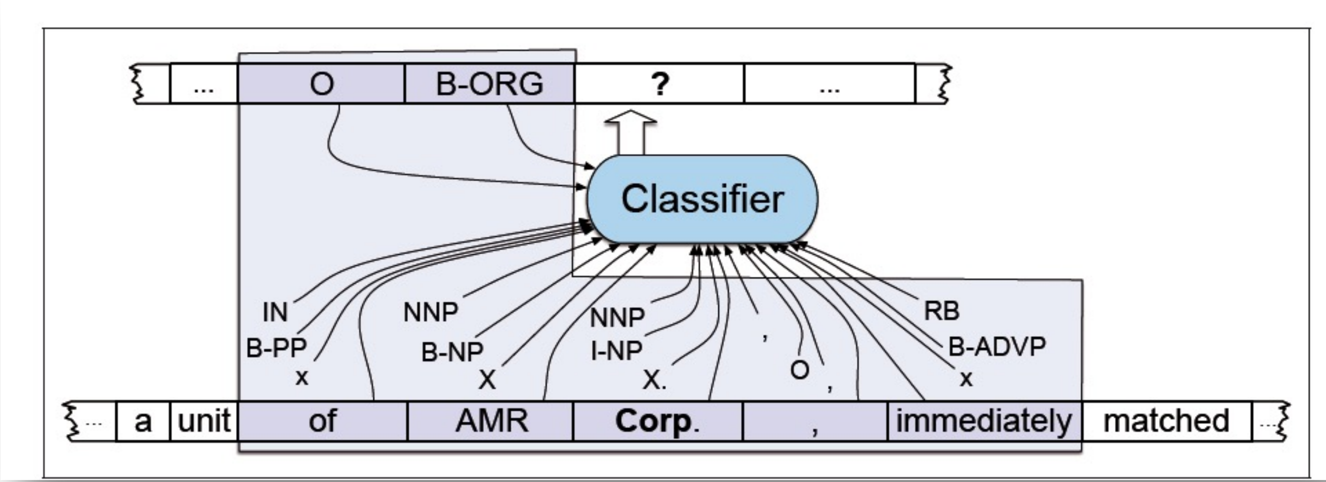
(Named) Entity Recognition using Machine Learning

- Other encoding schemas:
  - BE – model a span using its first (B) and its last token (E)
  - Usually applied for very long spans
- $3k + 1$ classes (but can be reduced as well)



NER as sequence classification

- We can then treat this task entirely the same manner as we did with POS-Tagging
  - Usually this is also carried out on a sentence level, using:
    - Maximum Entropy
    - MEMM
    - CRF



NER as sequence classification

- And use (very similar) features, such as:

1. Token at positions: 0,1,-1,-2,2,…
2. Concatenation of tokens in interval: [-2,2],[-2,1],[-2,0],[-1,2],…
3. Gazetteer contains token at: 0,1,-1,-2,2
4. POS-Tag at positions: 0,1,-1,-2,2,…
5. Concatenation of POS-Tags in interval: [-2,2],[-2,1],[-2,0],[-1,2],…
6. N-Grams (e.g. prefixes and suffixes)
7. Word-Shape (Uppercase, Lowercase, contains numbers, etc…)

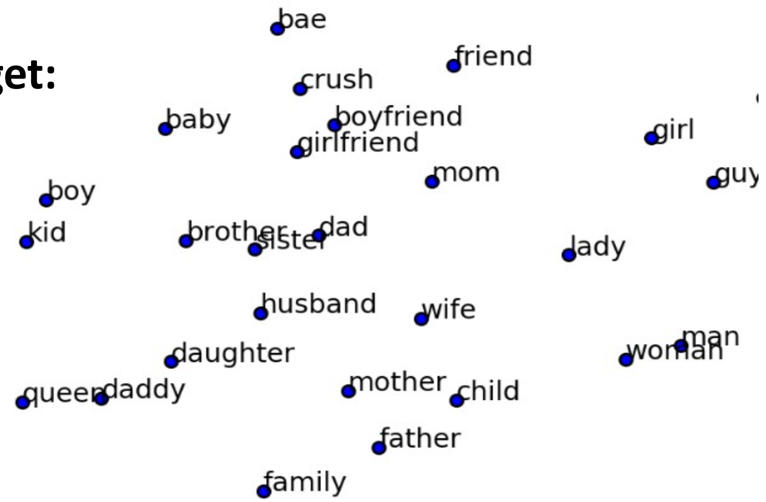
New:

Usage of cluster information

NER as sequence classification

- Usually way more unlabeled data than labeled data is available and we can and should use it!
- ➔ We are going to cluster the words of our large unlabeled data set into similar groups
  - → And use this information as additional features (e.g. “Cluster=Cluster12”)
  - → Called “Semantic Generalization” and part of **Semi-Supervised Learning**

One example cluster we are trying to get:

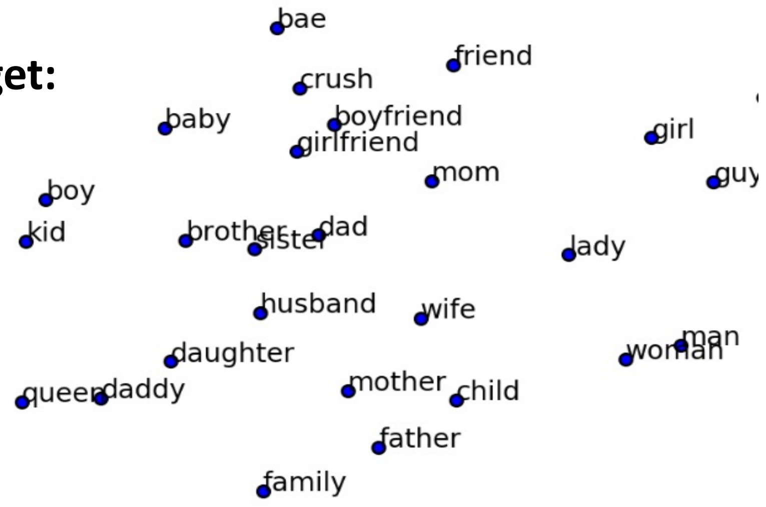


„Twitter Word2Vec“

NER as sequence classification

- Usually way more unlabeled data than labeled data is available and we can and should use it!
- ➔ We are going to cluster the words of our large unlabeled data set into similar groups
  - → And use this information as additional features (e.g. “Cluster=Cluster12”)
  - → Called “Semantic Generalization” and part of **Semi-Supervised Learning**

One example cluster we are trying to get:



„Twitter Word2Vec“

We will introduce the Brown Clustering and GloVe Embeddings for this purpose
➔ Stay tuned!