729G86/TDP030 Language Technology (VT2025)

Sequence Labelling

Marcel Bollmann

Department of Computer and Information Science (IDA)

LINKÖPING UNIVERSITY

Adapted from slides by Marco Kuhlmann.

Today's lecture

1. Introduction & Tasks
   - Part-of-Speech Tagging
   - Named Entity Recognition
   - Word Segmentation
2. Evaluation
   - POS Tagging
   - Span-Level Metrics for NER
3. Perceptrons
   - Features
   - Weights
   - Tagging
4. Outlook: Neural Networks

What is Sequence Labelling?

Introduction & Tasks

Sequence labelling

Previously...

Text classification is the task of categorizing **text documents** into predefined classes.

Definition

Sequence labelling is the task of annotating **each item in a sequence** with predefined labels.

- "Items in a sequence" can be e.g. words in a sentence

Part-of-speech tagging

The	*quick*	*brown*	*fox*	*jumped*	*over*	*the*	*lazy*	*dog*	.
DET	ADJ	ADJ	NOUN	VERB	ADP	DET	ADJ	NOUN	PUNCT

- A **part of speech** is a category of words that play similar roles within the syntactic structure of a sentence.
- Common parts of speech are **noun**, **verb**, or **adjective**.

Universal part-of-speech tagset

Tag	Category	Examples	Tag	Category	Examples
ADJ	adjective	*brown, old, big*	ADP	adposition	*of, in, to*
ADV	adverb	*quickly, very*	AUX	auxiliary	*has, is, will*
INTJ	interjection	*ouch, hello*	CCONJ	coord. conjunction	*and, or, but*
NOUN	noun	*dog, car, house*	SCONJ	subord. conjunction	*that, while*
PRON	pronoun	*you, her, theirs*	DET	determiner	*a, an, the*
PROPN	proper noun	*Maria, Berlin, IBM*	NUM	numeral	*42, fifteen*
VERB	verb	*writes, jumped*	PART	particle	*'s, not*
PUNCT	punctuation	*. , : ; ! ?*	SYM	symbol	*$ € :)*

Source and more details: Universal POS tags

Part-of-speech tagging

Definition

Part-of-speech (POS) tagging is the task of tagging each word in a sentence with its part of speech.

- There are many **different tagsets** for part-of-speech tagging.
  - Different levels of granularity, or tailored to different languages.
  - "Universal POS tagset" = universally applicable across languages, *not* "universally used"!
- This can be framed as a **supervised machine learning** problem.

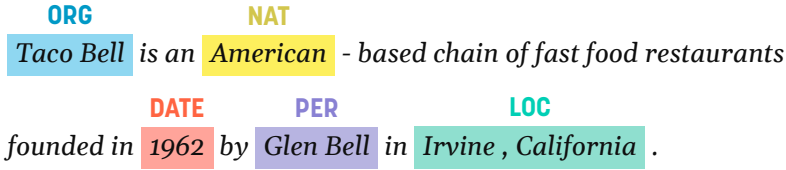
Ambiguity causes combinatorial explosion

The	*quick*	*brown*	*fox*	*jumped*	*over*	*the*	*lazy*	*dog*	.
DET	ADJ	ADJ	NOUN	VERB	ADP	DET	ADJ	NOUN	PUNCT
	ADV	NOUN	VERB		ADJ			VERB	
	NOUN	VERB			ADV				

Named entity recognition

Definition

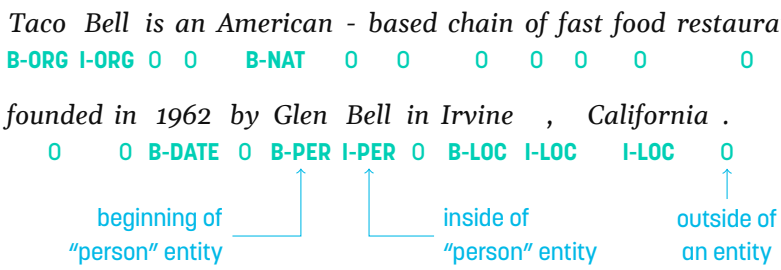
Named entity recognition (NER) is the task of identifying named entities and labelling them with their type.



Source: Wikipedia

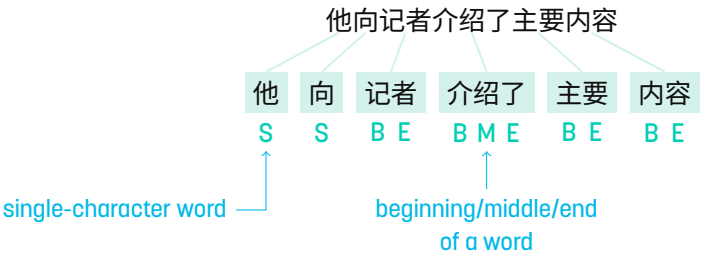
Named entity recognition as sequence labelling

- Word-level tags can encode both the **boundaries** and **types** of named entities.
- A common encoding scheme that implements this idea is **BIO notation**:



Chinese word segmentation

"He briefed reporters on the main contents"

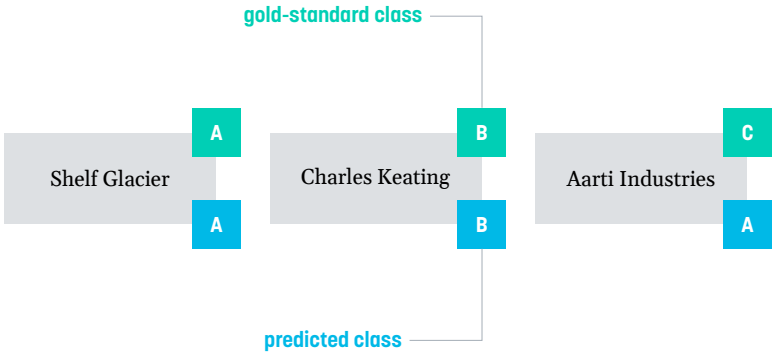


Important concepts

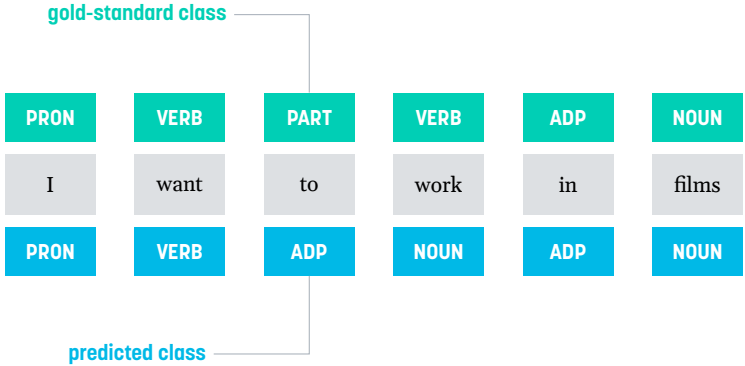
- part-of-speech (POS) tagging
- named entity recognition (NER)
- BIO notation

Evaluation of Sequence Labelling

Reminder: Evaluation of text classifiers



Evaluation of part-of-speech taggers



Accuracy

gold-standard \ predicted	DET	ADJ	NOUN	ADP	VERB
DET	923	0	0	0	1
ADJ	2	1255	132	1	5
NOUN	0	7	4499	1	18
ADP	0	0	0	2332	1
VERB	0	5	132	2	3436

$$\frac{12445}{12752} = 97.59\%$$

Precision with respect to NOUN

gold-standard \ predicted	DET	ADJ	NOUN	ADP	VERB
DET	923	0	0	0	1
ADJ	2	1255	132	1	5
NOUN	0	7	4499	1	18
ADP	0	0	0	2332	1
VERB	0	5	132	2	3436

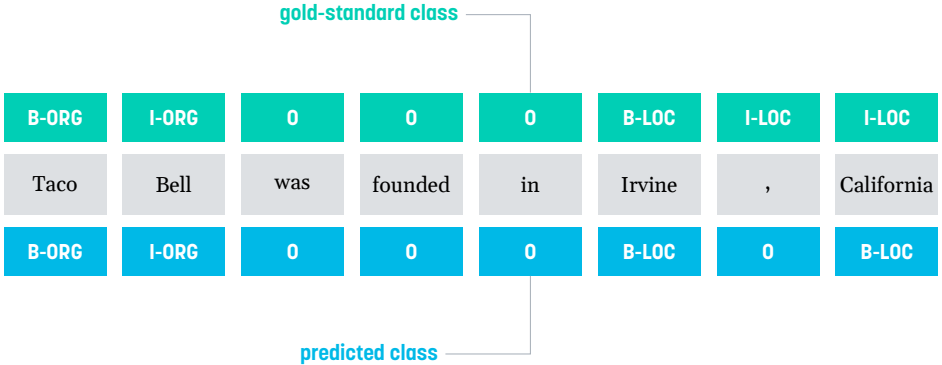
$$\frac{4499}{4763} = 94.46\%$$

Recall with respect to NOUN

gold-standard \ predicted	DET	ADJ	NOUN	ADP	VERB
DET	923	0	0	0	1
ADJ	2	1255	132	1	5
NOUN	0	7	4499	1	18
ADP	0	0	0	2332	1
VERB	0	5	132	2	3436

$$\frac{4499}{4525} = 99.43\%$$

Evaluation of named entity recognizers

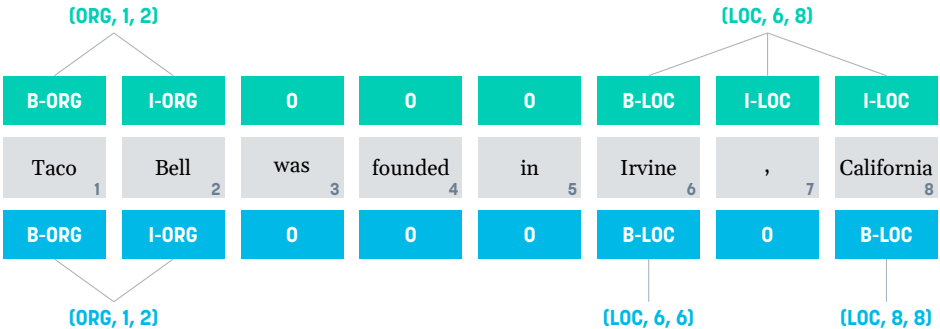


Problems with standard metrics for NER

B-ORG	I-ORG	O	O	O	B-LOC	I-LOC	I-LOC
Taco	Bell	was	founded	in	Irvine	,	California
B-ORG	I-ORG	O	O	O	B-LOC	O	B-LOC

- The O tag is the **most frequent** one, but also the one we **care least about**.
  → makes accuracy a bad choice here
- B-* and I-* tags always **belong together**.
  → doesn't make sense to compute precision/recall separately for them

Converting tags into spans

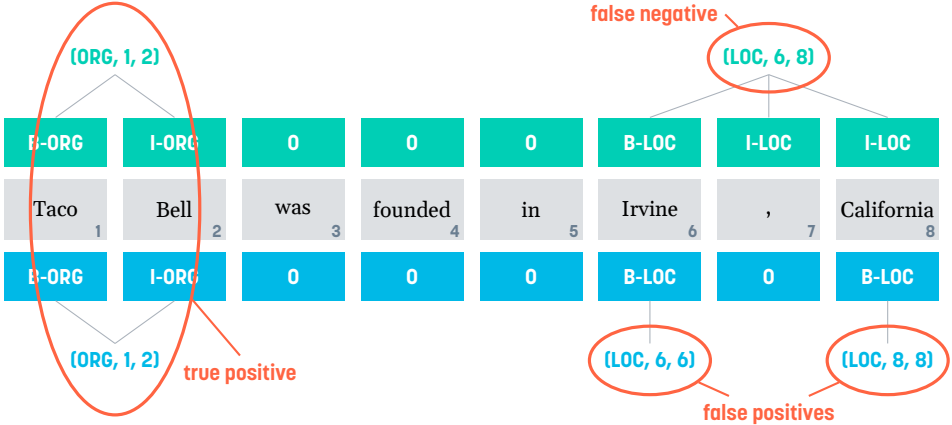


Reminder: Precision and recall with two classes

$$\text{precision} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false positives}}$$

$$\text{recall} = \frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}}$$

Span-level precision/recall for NER



$$\text{precision} = \frac{1}{3} = 33.33\%$$

$$\text{recall} = \frac{1}{2} = 50\%$$

Important concepts

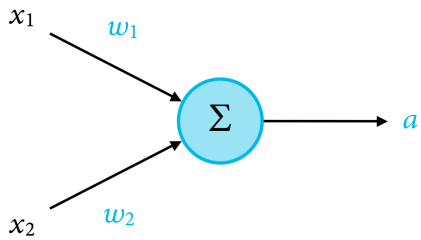
- accuracy, precision, recall, F1-score
- span-level precision and recall *(for NER)*

Sequence Labelling with Perceptrons

Sequence labelling as classification

- We can treat sequence labelling as a **classification problem**.
  - one classification per word in the sentence
  - similar to what we did for text classification!
- The **multi-class perceptron** is a very simple, **non-probabilistic** classifier.
  - a very simple type of neural network

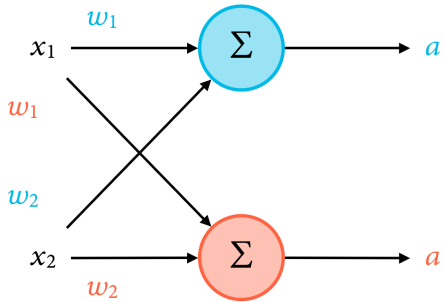
The classical perceptron



$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{xw} > 0 \\ 0 & \text{otherwise} \end{cases}$$

A perceptron is a **linear model** with a **decision rule**.

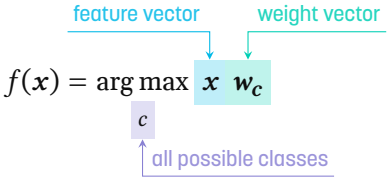
The multi-class perceptron



$$f(\boldsymbol{x}) = \underset{c}{\text{argmax}}\ \boldsymbol{x}\boldsymbol{w}_c$$

A multi-class perceptron uses a different decision rule to predict **multiple classes**.

The multi-class perceptron, formally



- The **feature vector** is the **input** – this is how the perceptron "sees" the data.
- The **weight vector** is a model **parameter** – this is what the perceptron "learns."

Feature vectors

feature vector

$$f(\boldsymbol{x}) = \arg\max_{c} \boldsymbol{x}\,\boldsymbol{w_c}$$

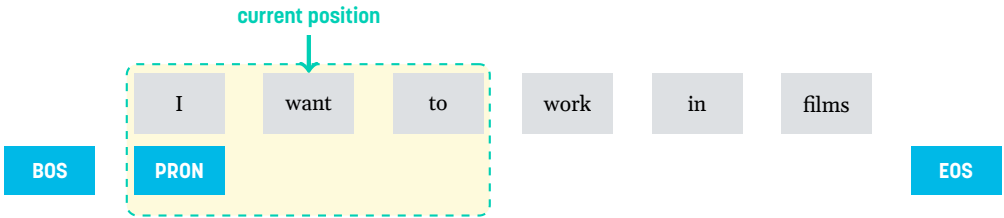
- Each **dimension** of $\boldsymbol{x}$ corresponds to one **feature**.
  - *Example:* “the word form of the current token is ‘films’”
- Which **features to extract** from the data is a decision that we have to make!

What can go in a feature vector?

- Features can be any information that can be derived from **the input sentence** or any **previously predicted labels**.
  - *Example:* the word form of the current token
  - *Example:* the part-of-speech tag of the previous token
- This means we can **look further back** or even **look ahead**.
- At the same time, using *too* much information can lead to problems.
  - efficiency, data sparseness

Feature windows

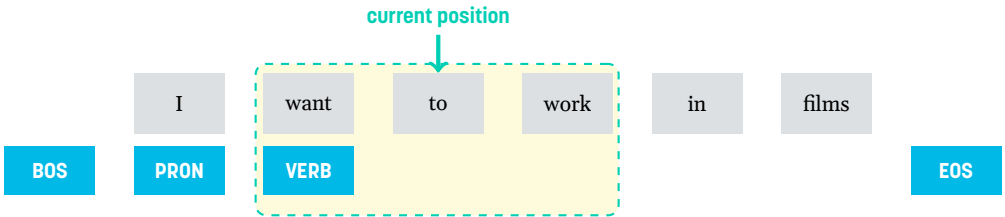
- A compromise is to define a limited **feature window**, for example of 1:



- Here we can use the previous word, current word, next word, and the previous tag.

Feature windows

- The feature window **moves forward** during tagging:



Examples of features in part-of-speech tagging

- word form (lowercased) of the current token
- word form of the preceding/next token
- capitalization of the current token (upper, lower, n/a)
- type of the current token (digits, letters, symbols)
- prefixes and suffixes of the current token (of various length)
- whether the current token is hyphenated
- whether the token is first or last in the sentence
- various combinations of the other features

Source: Östling (2013)

Chocolate box challenge!

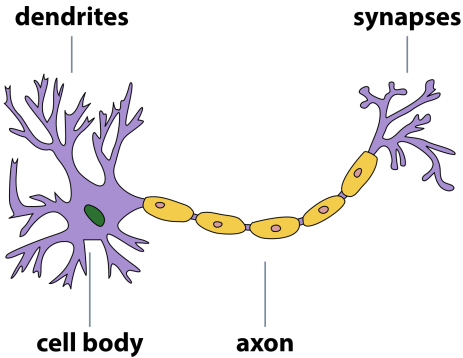
Weight vectors

weight vector

$$f(\boldsymbol{x}) = \arg\max_{c} \boldsymbol{x}\, \boldsymbol{w_c}$$

- The **dimensions** of $\boldsymbol{w_c}$ correspond to the **importance of that feature** for the class c.
- If $\boldsymbol{w_{c,i}} > 0$, the feature $\boldsymbol{x_i}$ **does belong** to class c.
- If $\boldsymbol{w_{c,i}} < 0$, the feature $\boldsymbol{x_i}$ **does *not* belong** to class c.
  - We assume that feature values x_i are always non-negative.

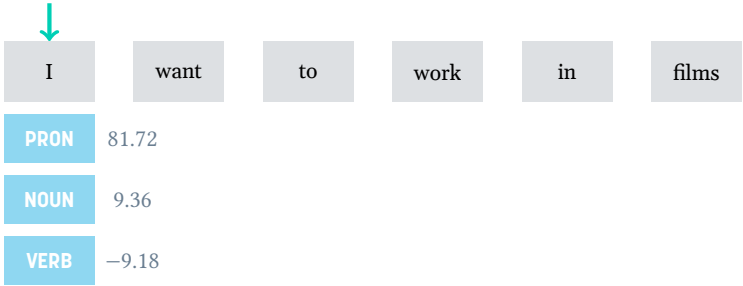
Inspiration from neurobiology



- Features whose weights are positive **increase the activation** of the neuron.
- Features whose weights are negative **decrease the activation** of the neuron.
- Features whose weights are zero **do not contribute** anything to the activation of the neuron.

Part-of-speech tagging with a perceptron

- We tag our sentence **from left to right**, picking the highest-scoring tag.

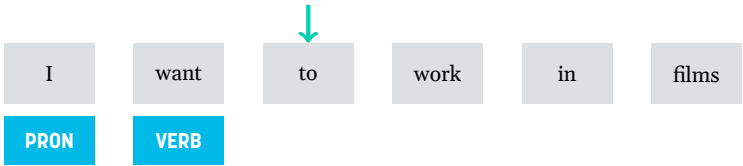


Part-of-speech tagging with a perceptron

- Note that the scores are **not probabilities** anymore!

Part-of-speech tagging with a perceptron

- Continue until the end of the sequence.



Important concepts

- (multi-class) perceptron
- feature vector
- feature window

Outlook: Neural Networks

State of the art

- Like most language technology applications, the current state-of-the-art models for sequence labelling rely on **artificial neural networks**.

- **Perceptrons** are the "simplest" form of neural networks.
  - ANNs add a **non-linear activation function** – e.g. $\tanh(\boldsymbol{x}\boldsymbol{w}_c)$ instead of just $\boldsymbol{x}\boldsymbol{w}_c$.
  - ANNs **chain multiple "neurons" together**, so that the output of one becomes the input of another.


LINKÖPING
UNIVERSITY