Tasks, resources and evaluation PART-1

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

Overview

- Evaluation goals and process
- Evaluation issues, methodologies and metrics
- Distribution of resources: consortia, associations and campaigns

Evaluation

Only by applying some form of evaluation
can we prove that

an NLP tool has the expected ability to perform the
task assigned to it

and/or that a resource contains the expected
knowledge.

Evaluation

The Evaluation of a resource can be only performed by training a system on it. **A resource can only be evaluated indirectly.**

I can internally observe resource for a given language, i.e. make a qualitative and quantitative analysis of its content, but this is not enough to decide if its content is what I need for dealing with that language.

It is by training a statistical model on the resource that I can understand whether it really contains what is needed to treat the language it represents.

Evaluation

The Evaluation is an inherent activity of AI and NLP.

It usually consists of a **comparison between humans and machines**, and therefore focuses on both **knowledge** (contained in resources) and **skills** (exhibited by a system using the resources).

The reference point of evaluation is generally human skill and knowledge, which is expected of a machine capable of simulating human behaviour in a linguistic task.

Evaluation

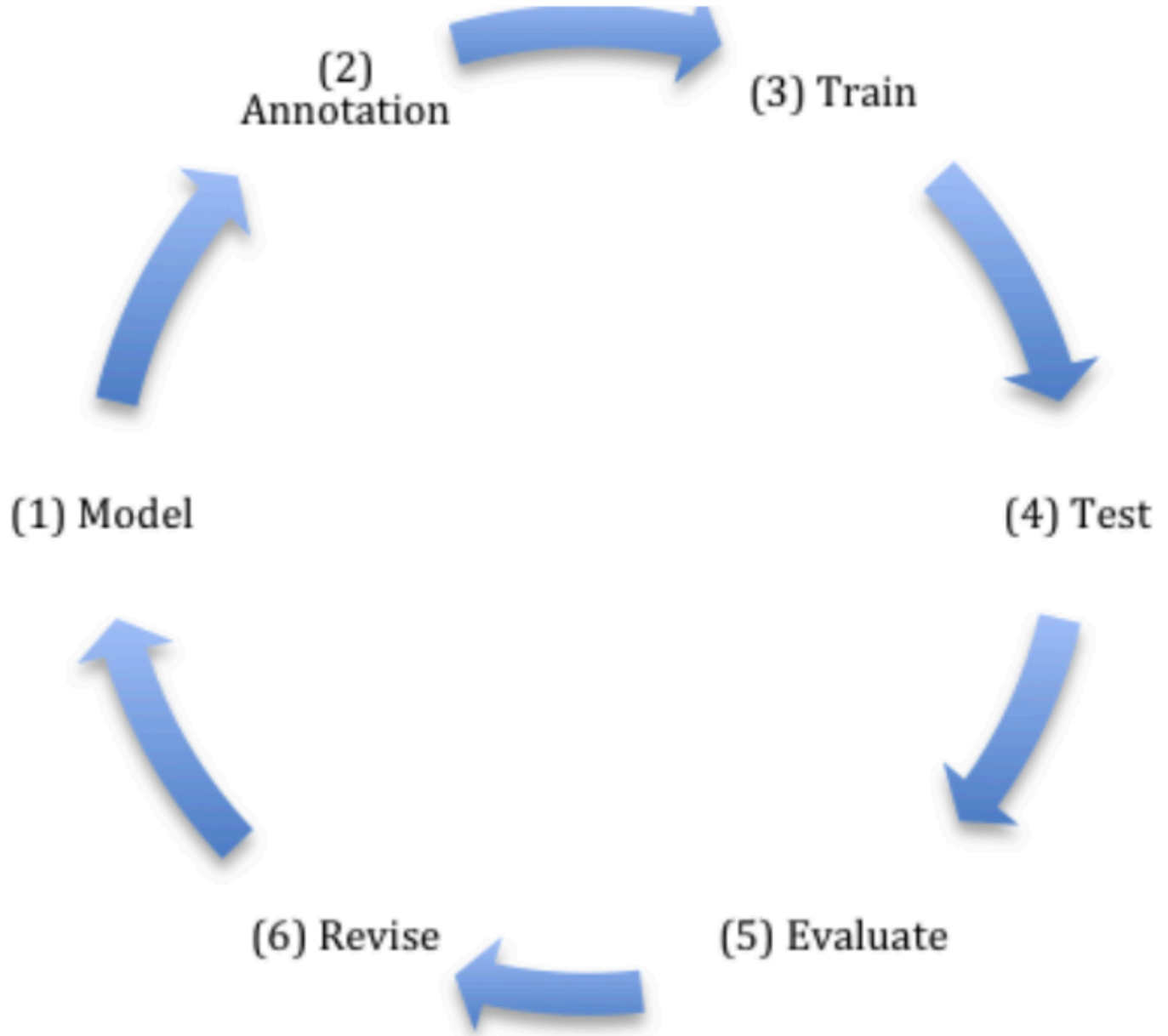
Evaluation is applied routinely **during the development** of a resource to have precise information about its quality
> this is why MATTER is a ***cycle*** and not a simple sequence of steps.

It is applied in **comparative** settings to have knowledge of the differences in performance that can be achieved by different analysis tools based on different resources
> this is the motivation for which are organised comparative evaluation exercises (i.e. ***shared tasks*** within evaluation campaigns).

Gold standard

According to the **MATTER cycle**:

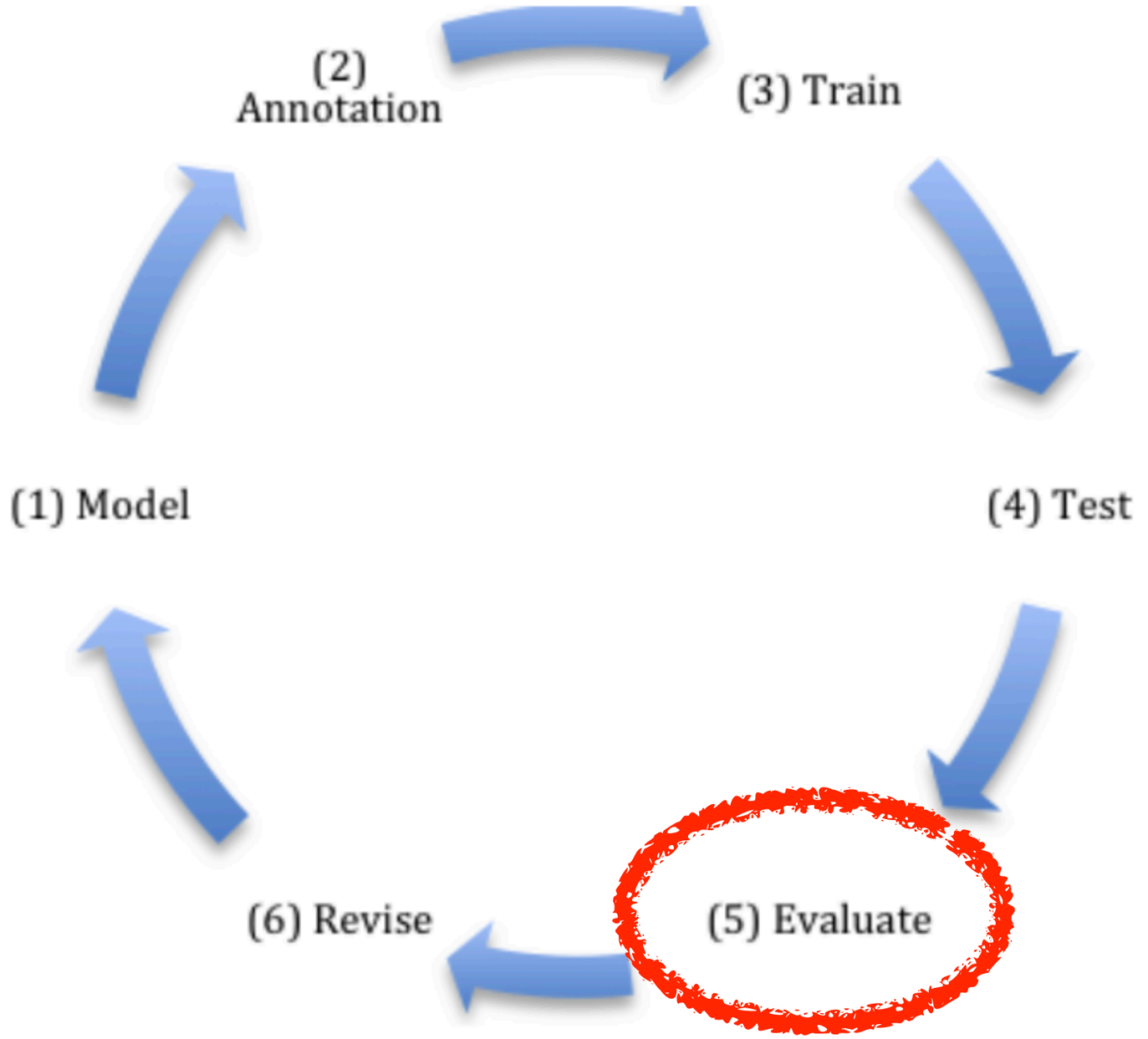
- 1+2 we develop the resource (gold standard = training set + test set)
- 3 we train a statistical model on the **training set**
- 4 we test the model on the **test set**
- 5 we evaluate the results



Gold standard

According to the **MATTER cycle**:

- 1+2 we develop the resource (gold standard = training set + test set)
- 3 we train a statistical model on the **training set**
- 4 we test the model on the **test set**
- 5 we evaluate the results



Gold standard

How can we prove that a model is good for the language we are dealing with?

We provide a set of data, we ask human judges to carefully annotate them (possibly using automatic tools and then correcting their outputs)

We train the machine with a portion (80%) of the annotated data

We evaluate the results with another portion (20%) of the annotated data. **Are the results the same as the data annotated by human judges?**

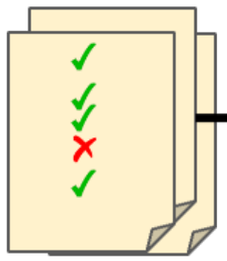
Gold standard

The training set is the focus of the learning process.

The test set is the focus of the evaluation process.

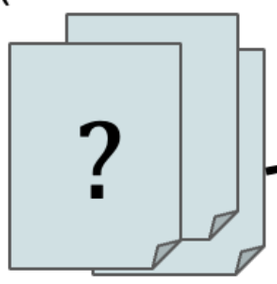
They both represent the knowledge of the community of speakers about the represented language.

Training set
with Gold standard Annotation

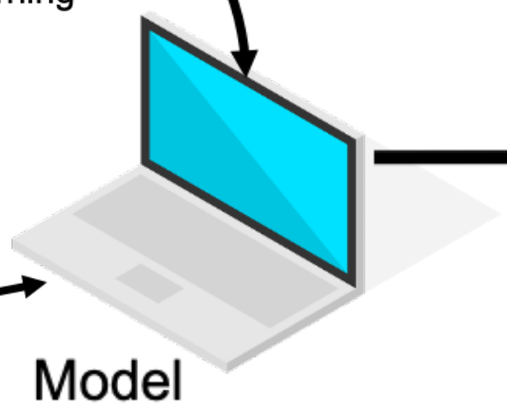


Machine Learning

Novel data
(to be annotated)



Test set
Without Annotation



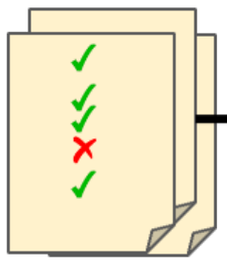
Model

Novel data
(annotated)

Test set
With automatic Annotation

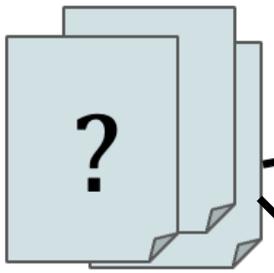
Evaluation process

Training set
with Gold standard Annotation

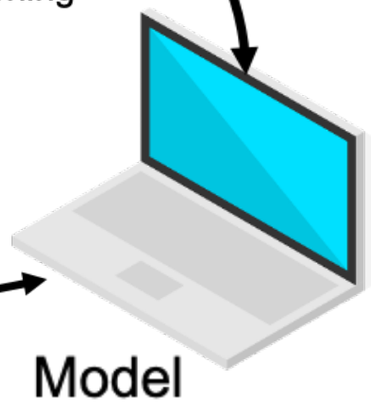


Machine Learning

Novel data
(to be annotated)

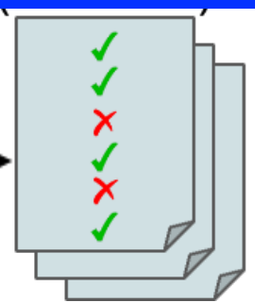


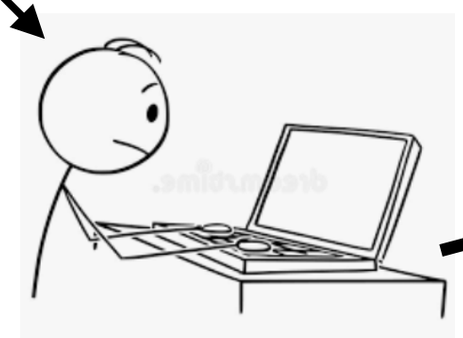
Test set
Without Annotation

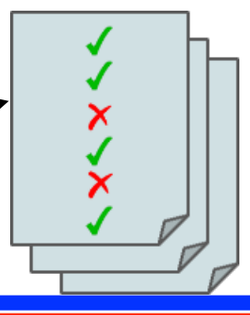


Model

Test set
With automatic Annotation

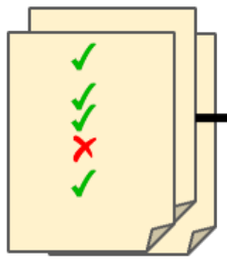






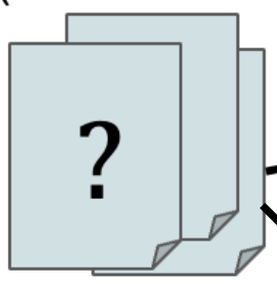
Test set
With gold standard Annotation

Training set
with Gold standard Annotation

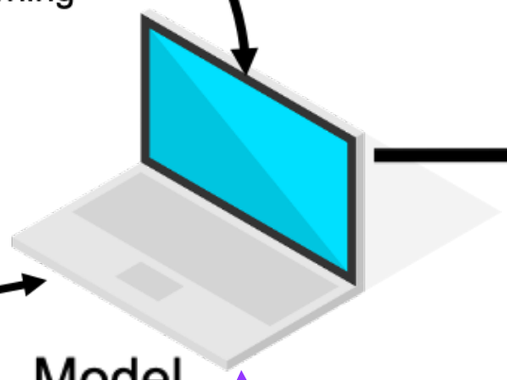


Machine Learning

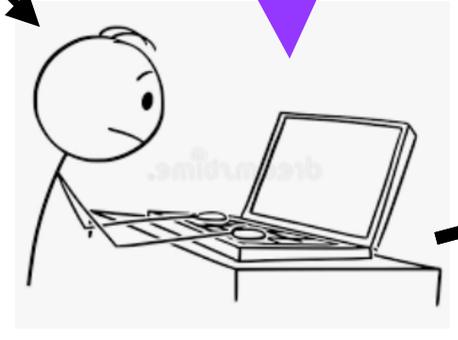
Novel data
(to be annotated)



Test set
Without Annotation

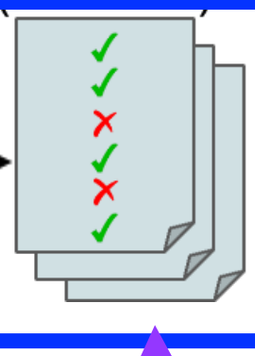


Model

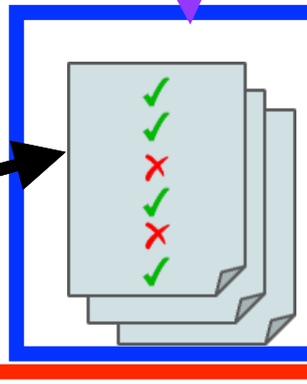


Evaluation process

Test set
With automatic Annotation



Evaluation outcome



Test set
With gold standard Annotation

Evaluation issues

- Should the model correctly classify **all the data** in that test set?

- Can we accept a model that classifies only a part of the test set? Which part?

- Are all errors detected in the results of equal magnitude? Or are there more and less serious errors?

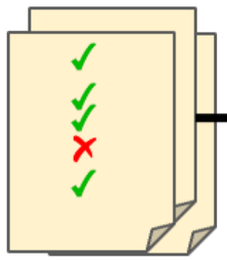
Baseline

In some cases, **a comparison between human** capabilities / knowledge **and machine** systems / resources in terms of performance **is not possible or meaningful**.

Or the **evaluation** has less ambitious goals in terms of skills and **does not** really **refer to human performance** but to the results previously achieved by other systems, the so-called **baseline**.

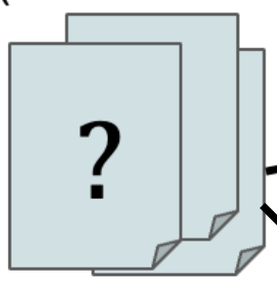
A baseline represents a reasonable minimum expected system outcome.

Training set
with Gold standard Annotation

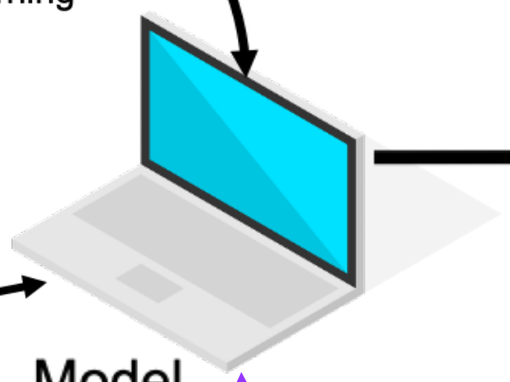


Machine Learning

Novel data
(to be annotated)



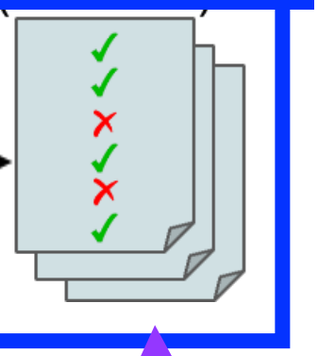
Test set
Without Annotation



Model

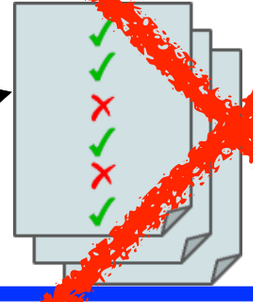
Evaluation process

Test set
With automatic Annotation

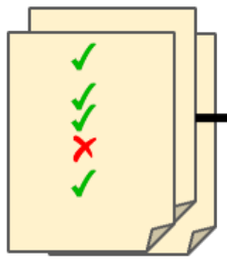


Evaluation outcome

Test set
With gold standard Annotation

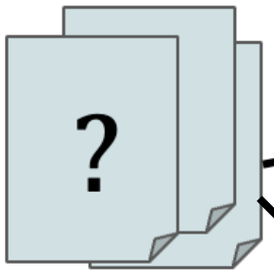


Training set
with Gold standard Annotation

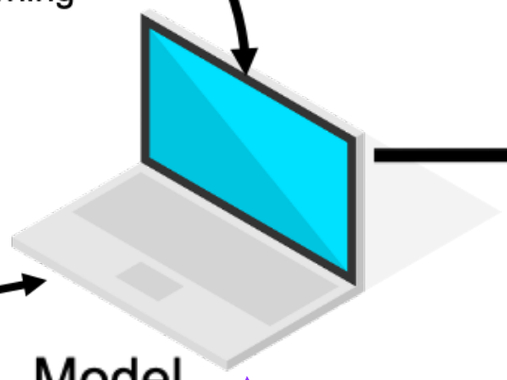


Machine Learning

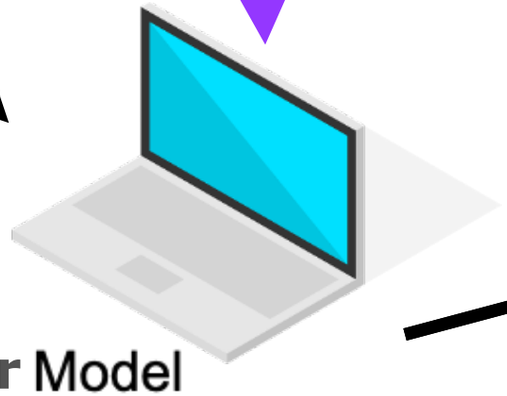
Novel data
(to be annotated)



Test set
Without Annotation



Model

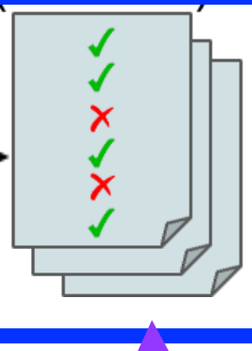


Other Model

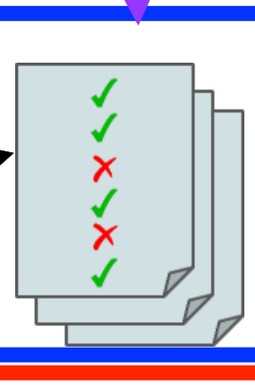
BASELINE

Evaluation process

Test set
With automatic Annotation



Evaluation outcome



Test set
With automatic Annotation

Baseline

For example, machines are better than humans at recognizing the genre of a speaker or writer (this can be a subtask of a task called *author profiling*). For this task it is important to use baselines.

This is a linguistic task that humans rarely attempt to perform, because we can detect genre by simply observing the speaker or because it is not an information about the author we want to detect when we read a written document.
This task can be important in the context of detecting people who impersonate others for fraudulent purposes (such as terrorists or paedophiles).

Baseline

For other tasks that are known as too difficult for a machine, the baseline represents an acceptable result to be achieved.

This is the case of paraphrasing or summarising, in which machine performance is still meaningful lower than human.

Evaluation metrics

Only if we apply some evaluation metrics
we can make **evaluation objective**
and we can later **compare** results achieved
in different settings.

Less and more specific metrics exist that are designed
for several tasks or for a single task.

Evaluation metrics

Linguists are used to performing subjective analyses, but they can be not appropriate for NLP.

Comparison of machines with human skills and knowledge is done in NLP with the help of **specific metrics and methods** that make possible an objective assessment of system results.

Objective evaluation, in turn, can be useful to **improve system performance** on a given task, as it **provides insight into system and resource limitations**.

Metrics

Given the test set, we can count how many times the output provided by the model corresponds to that in the gold standard test set.

Different metrics help us in answering to questions as

How can we count?
What can be count?

For example, in PoS tagging we can count how many words are correctly classified by our model.

Accuracy

Given the gold standard test set, we can count how many times the output delivered by the model is equal to that in the gold standard test set.

This measure is called **ACCURACY** and is equal to the number of correct outputs divided by the number of required outputs.

$$\frac{correctResults}{requiredResults}$$

PoS tagging

Unannotated TEST SET	Model output = automatic annotation	GOLD standard test set
The	DET	DET
cat	NOUN	NOUN
run	NOUN	VERB
in	PREP	PREP
the	DET	DET
garden	NOUN	NOUN
Mary	PrNOUN	PrNOUN
sleeps	VERB	VERB
in	PREP	PREP
the	DET	DET
sun	ADV	NOUN
…	…	…

PoS tagging

TEST SET	Model output	GOLD	Evaluation
The	DET	DET	I
cat	NOUN	NOUN	I
run	NOUN	VERB	0
in	PREP	PREP	I
the	DET	DET	I
garden	NOUN	NOUN	I
Mary	PrNOUN	PrNOUN	I
sleeps	VERB	VERB	I
in	PREP	PREP	I
the	DET	DET	I
sun	ADV	NOUN	0
…	…	…	…

PoS tagging

TEST SET	Model output	GOLD	Evaluation
The	DET	DET	I
cat	NOUN	NOUN	I
run	NOUN	VERB	0
in	PREP	PREP	I
the	DET	DET	I
garden	NOUN	NOUN	I
Mary	PrNOUN	PrNOUN	I
sleeps	VERB	VERB	I
in	PREP	PREP	I
the	DET	DET	I
sun	ADV	NOUN	0
…	…	…	…

Accuracy = 9 : 11 = 0,8181…

PoS tagging

TEST SET	Model output	GOLD	Evaluation
The	DET	DET	I
cat	NOUN	NOUN	I
run	NOUN	VERB	0
in	PREP	PREP	I
the	DET	DET	I
garden	NOUN	NOUN	I
Mary	PrNOUN	PrNOUN	I
sleeps	VERB	VERB	I
in	PREP	PREP	I
the	DET	DET	I
sun	ADV	NOUN	0
…	…	…	…

Accuracy = 9 : 11 = 0,8181…

which expressed in percentage is ~81%

Accuracy

Accuracy **is not the best measure** in some cases.

For example:
- the task consists in classifying words according to only two classes (such as VERB and nonVERB)
- *the test set will probably include a high percentage of nonVERB*
- a model that takes into account this information can consist in simply assigning always the same class nonVERB to all the instances to be annotated
- result: the model will be quite accurate also without encoding any interesting knowledge!

***Partial* PoS tagging** (with 2 tags only)

TEST SET	Model output	GOLD	Evaluation
The	nonVERB	nonVERB	1
cat	nonVERB	nonVERB	1
run	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
garden	nonVERB	nonVERB	1
Mary	nonVERB	nonVERB	1
sleeps	nonVERB	VERB	0
in	nonVERB	nonVERB	1
the	nonVERB	nonVERB	1
sun	nonVERB	nonVERB	1
…	…	…	…

Accuracy = 9 : 11 = ~81%

Baseline

It is also **to avoid models that do not really encode interesting knowledge**, that it is important to establish a **baseline** to be used as a reference in the evaluation process.

As a baseline we can take the result you get completely at random (accuracy 50%) as in rolling dice, or by always choosing the class you know to be the most frequent.

In this way, we are sure that only if the model encodes some linguistic knowledge it must give a result that is above the baseline.

Beyond accuracy

Another solution for better evaluating system performance is to **separately observe** the model on the different **categories to be classified** for the task.

We focus in turn on a single category only, by non considering the other ones and evaluating the system against it alone.

TEST SET	Model output	GOLD
The	VERB	nonVERB
cat	nonVERB	nonVERB
run	nonVERB	VERB
in	nonVERB	nonVERB
the	nonVERB	nonVERB
garden	VERB	nonVERB
Mary	nonVERB	nonVERB
sleeps	VERB	VERB
in	nonVERB	nonVERB
the	nonVERB	nonVERB
sun	VERB	nonVERB
…	…	…

Accuracy for VERB = 1 : 2 = 50%

Accuracy for nonVERB = 6 : 9 = ~ 66%

Beyond accuracy

By separately observing each class to be categorised by the model, we can provide more precise evaluations of its performance.

For example, focusing on the class nonVERB, we can see that our model:

- correctly assigns nonVERB to some words (*cat, in, the, Mary, in, the*), which are the **TRUE POSITIVE**s in this classification

- incorrectly assigns nonVERB to some other word (*run*), which are the **FALSE POSITIVE**s.