Development of resources

Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2023-24

Cristina Bosco

Overview

- Annotations and tasks

- Main steps in the development of a linguistic resource

Annotation and tasks

The annotation of a linguistic resource can be quite complex and involve a whole team.

There are some main steps in the development of a resource

They may depend on the kind of resource, on the purpose for which it is created and the specific task in which it will be used

Annotation: a definition

Annotation consists of adding linguistic information to the pure text, i.e. making the linguistic knowledge implicit in data explicit.

The process by which the data required for training linguistic models is generated is usually simply called annotation.
In practice, however, this term conceals several steps.

Several different aspects of linguistic data can be annotated.
Although the annotation process may vary depending on the aspects we want to annotate, there are some important commonalities between all annotation processes.

Annotation: a definition

As examples of annotation, we have seen PoS tagging and parsing. The former makes the morphological linguistic knowledge implicit in the data explicit, while the latter does the same for syntactic knowledge.

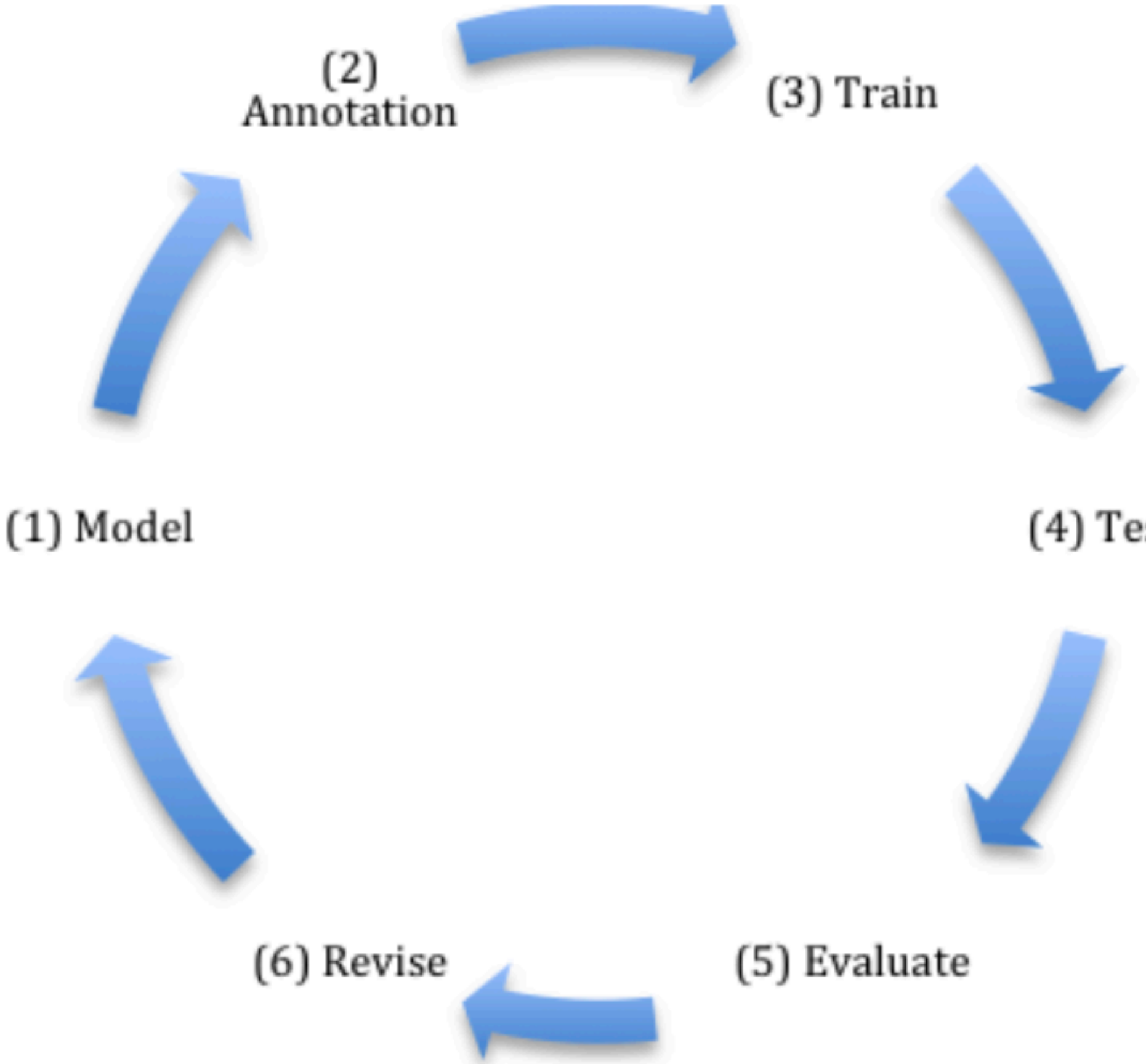
The aspects that differ in PoS tagging compared to parsing include the form of input and output, but also the algorithms used: PoS tagging analyses each word without context, whereas parsing attempts to integrate each incoming word into the syntactic structure (from left to right).

The MATTER cycle

The annotation process of a linguistic resource can be formally defined using the cycle MATTER[1] whose name is the acronym of the names of the six steps that compose the cycle:

MODEL
ANNOTATE
TRAIN
TEST
EVALUATE
REVISE

[1] James Pustejovsky and Amber Stubbs.(2012) Natural Language Annotation for Machine Learning. O'REILLY.


(1) Model
(2) Annotation
(3) Train
(4) Test
(5) Evaluate
(6) Revise

Step 1: Model

The first step in MATTER is dedicated to defining a conceptual framework for the phenomenon to be described: the phenomenon is carefully analyzed by the team developing the resource, taking into account theoretical studies about it and empirically observing the available data.

The results of this phase, that is a sort of **conceptual modelling** (not to be confused with the statistical model), are:

- a careful **description** of the linguistic characteristics of the phenomenon
- a preliminary definition of the **annotation scheme** to be used in the annotation
- a set of **guidelines** to be used by the annotators.

Step 1: Model

The **annotation scheme** is the set of labels to be used in labelling data.

There are one or more valid values for each label.

The labels can be clustered in different groups according to their meaning or organised hierarchically when the annotation can be applied to data as a cascade process.

Step 1: Model

The **guidelines** are the set of rules about the association phenomenon / label that must be followed by the annotators during the annotation process.

Following these guidelines each independent annotator can provide consistent data, that is data in which the same phenomenon is always annotated exactly in the same way.

The guidelines include in particular several examples in order to provide all the explanations about how to annotate specific (and also controversial) cases.

Model: an example

For example, in the Part of Speech tagging task we have performed in the exercises:

- the **conceptual model** is the description of the linguistic characteristics of the morphology we discussed
- the **annotation scheme** to be used in the annotation is that provided by the Penn Treebank project

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	*and, but, or*	NNP	proper noun, sing.	*IBM*	TO	"to"	*to*
CD	cardinal number	*one, two*	NNPS	proper noun, plu.	*Carolinas*	UH	interjection	*ah, oops*
DT	determiner	*a, the*	NNS	noun, plural	*llamas*	VB	verb base	*eat*
EX	existential 'there'	*there*	PDT	predeterminer	*all, both*	VBD	verb past tense	*ate*
FW	foreign word	*mea culpa*	POS	possessive ending	*'s*	VBG	verb gerund	*eating*
IN	preposition/ subordin-conj	*of, in, by*	PRP	personal pronoun	*I, you, he*	VBN	verb past participle	*eaten*
JJ	adjective	*yellow*	PRP$	possess. pronoun	*your, one's*	VBP	verb non-3sg-pr	*eat*
JJR	comparative adj	*bigger*	RB	adverb	*quickly*	VBZ	verb 3sg pres	*eats*
JJS	superlative adj	*wildest*	RBR	comparative adv	*faster*	WDT	wh-determ.	*which, that*
LS	list item marker	*1, 2, One*	RBS	superlatv. adv	*fastest*	WP	wh-pronoun	*what, who*
MD	modal	*can, should*	RP	particle	*up, off*	WP$	wh-possess.	*whose*
NN	sing or mass noun	*llama*	SYM	symbol	+,%, &	WRB	wh-adverb	*how, where*

- the set of **guidelines** are the official ones published for this resource (*Guidelines for the Part of Speech tagging of the Penn Treebank*)

Model: an example

For example, in the Part of Speech tagging task we have performed in the exercises:

- the labels can be clustered e.g. in those for nouns versus those for verbs
- the labels we used cannot be organised hierarchically
- there is a single value for each label

Step 2: Annotate

The second step in MATTER is dedicated to the application of the model to the data to be annotated and includes some subtasks where are addressed:

- the selection of **data** to be annotated
- the selection of **tools** to be used in the annotation
- **Annotators training** and management of the annotation team
- Evaluation of the annotated data.

Step 2: Annotate

The **selection of data** to be annotated is based on the interest of the research team and on the availability and accessibility of data.

For example:

- the research team is especially interested in a specific language and time-slot
- data corresponding to this interest are available for free on the web and can be collected and used
- data are not available because they are copyrighted or protected by a licence
- data must be collected using specific procedures (such as APIs from a social media platform)
- …

Step 2: Annotate

The **selection of tools** to be used in the annotation is usually based on the availability of platforms and their adequateness for the specific task.

There are several platforms designed for performing annotation tasks that can be customised (such as LabelStudio).

For specific tasks part of the annotation work can be done automatically and then manually corrected.

In some cases, in particular for annotations at sentence level, also simple spreadsheets can be used.

Step 2: Annotate

The **annotators training** is an important procedure that in more or less formal way allow to generate in the human judges involved in the annotation the awareness of the phenomena they have to analyse.

Usually a pilot annotation task is organised in which all the annotators are asked to annotate the same set of data and to discuss all together their results. This means that all the members of the annotation team annotate a small set of data and then discuss the differences in their annotations.

Step 2: Annotate

The **evaluation of the annotation results** allows to see whether the annotators are generating consistent annotated data.

When inconsistencies are detected in the results the causes must be investigated:

- the annotation scheme must be analysed > there are categories that are not clearly distinguishable?
- the guidelines must be extended and made more clear > there are phenomena and cases that are not included?
- the annotators must be more trained > there is some annotator that is not skilled enough about the guidelines?
- the annotated data must be carefully revised

Step 2: Annotate

A special role in the evaluation of the results is played by the calculation of **annotation disagreement** among the annotators.

The comparison of the annotations provided by different annotators shows whether some annotators is failing the assignment of categories to instances to be annotated.

Step 2: Annotate

The agreement of the annotators is crucial in tasks where there is only one correct answer, such as Part of Speech tagging. In cases in which there are more possible answers, it can be the object of further analysis.

Only if the data are internally correct and consistent they convey a conceptual model to the system that will be later trained on them.

If some occurrence of the word “dog" is classified as NOUN and some other as VERB, the conceptual model is not correct or its application to the data, therefore the statistical model will be confused.

Step 2: Annotate

For tasks where there is only a single correct answer for each item to be annotated, a **gold standard dataset** is released when an agreement is reached between the annotators by correcting the errors of each annotator and reaching a common conceptual model of the phenomena to be annotated.

A gold standard dataset, that is a carefully annotated and validated dataset after resolving possible disagreements, is crucial for the following two phases of the MATTER cycle.
It is a kind of representation of the **linguistic knowledge shared by the speaker community** about the annotated phenomena, i.e. in practice a complete representation of the conceptual model of the data.

Annotate: an example

For example in the Part of Speech tagging task we have performed in some exercises:

- we used a simple text editor for the annotation from scratch of some sentences > this is not the best solution for very fine-grained annotation because this procedure is to prone to errors

- we applied check and correction on pre-annotated data > this is the procedure applied in the development of the Penn Treebank

- we discussed together the results provided by the annotation team members and the guidelines > we didn't detected disagreements … because the task is very easy (…or the annotators very skilled 😁!!!)

Step 3 and 4: Train & Test

The third and fourth steps in MATTER are dedicated to learn the linguistic knowledge made available in the annotated data and to test the data.

They include the splitting of the data in two sets:

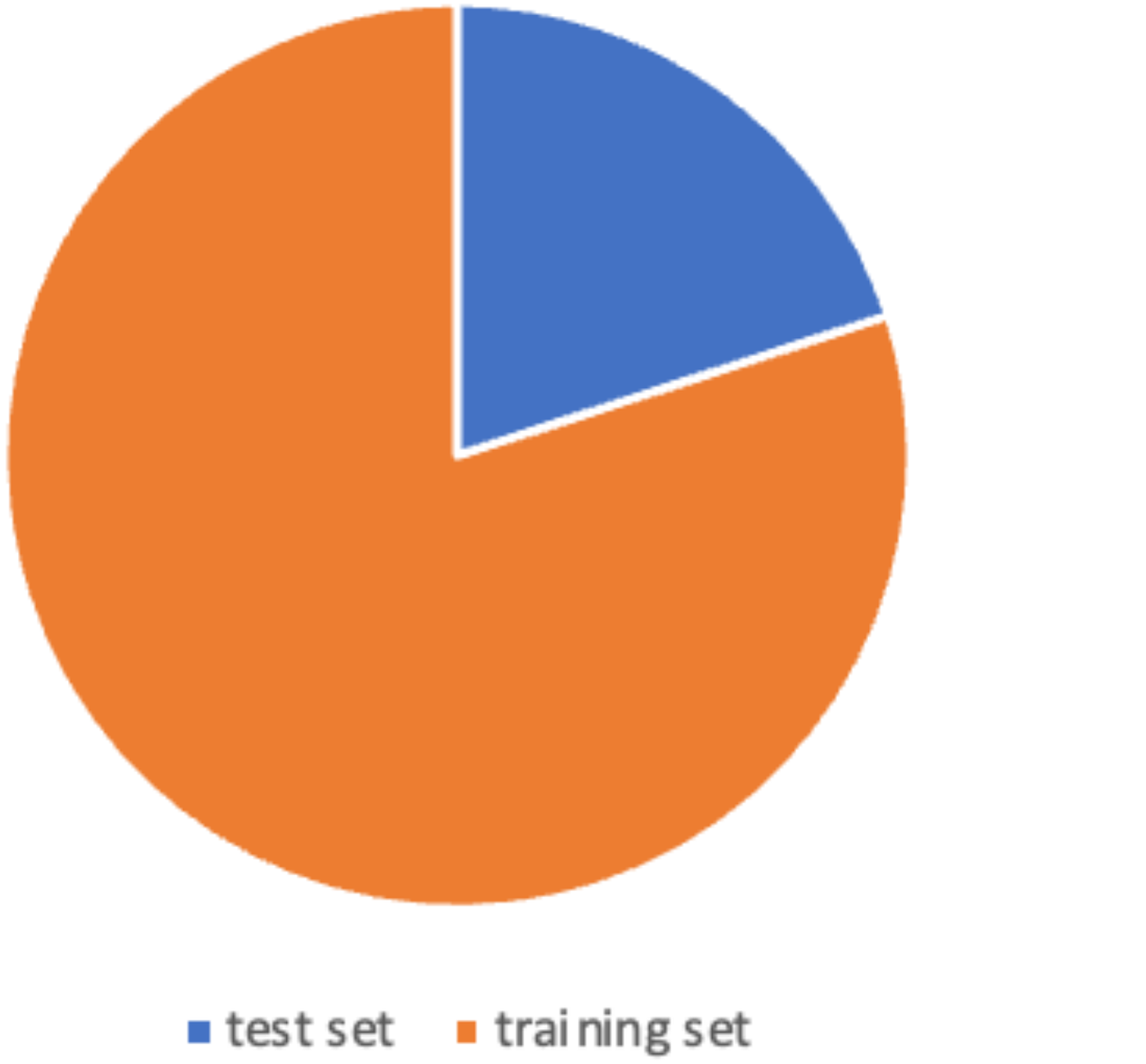
- the **training set** (or gold standard training set) on which the statistical model is built applying machine learning

- the **test set** (or gold standard test set) on which the statical model is tested.

Gold = Train + Test

The gold standard corpus is divided into a test set and a training set.

These parts of the corpus are then used in different ways.



Step 3 and 4: Train & Test

The **split between training set and test set** is usually 80% vs. 20%, whereby it is generally assumed that the more data is used, the more knowledge is learned and the best results are achieved.

Sometimes data that has only been automatically annotated (without human control) can also be used to train a model together with the gold standard training set data.
This is called **silver standard training data**.
Their quality is lower than that of the gold data training set, but they can be available in greater quantity and in less time, as they do not require the work of human annotators

Step 3 and 4: Train & Test

The **test set** must be provided in two versions:

- Unannotated > this version will be given to the system to perform its task, i.e. apply automatically the annotation

- Annotated as a gold standard > this version will be compared with the output of the system.

The test set is always and necessarily manually annotated (or checked) because it represents the reference for the evaluation of the results provided by the machine.

Step 3 and 4: Train & Test

The **pipeline** is as follows:

- Split of the annotated data in training and test set
- Application of machine learning to the training set to built the statistical model
- Use of the statistical model on the unannotated test set > output the test set automatically annotated
- Comparison of the gold standard (annotated) test set with the output automatically annotated > Evaluation (phase 5 of MATTER)

Step 5: Evaluate

The fifth step in MATTER id dedicated to the evaluation of the results that can be achieved by a machine using the annotated data, those included in the gold standard training set.

If the output of the machine applied on the (unannotated) test set is not the same as in the annotated gold standard test set, the evaluation can detect the limits of the annotated data to see:

- If some phenomenon is not represented or underrepresented in the data
- If some categories are not annotated in the right way, with inconsistencies
- If the data provided for training are not enough
- …

Step 5: Evaluate

The evaluation can be based on different systems using different statistical models:

- Training different systems on the same gold standard training set and testing them on the same test set we can see if the issues raised in the evaluation depend on the training data or on the system (and statistical model).

To ensure effective comparability of the systems and correct evaluation of the annotated data, it is important that the evaluation is based on the same data for all systems. This is a methodology applied in evaluation campaigns.

Step 6: Revise

The last step in MATTER is dedicated to the revision of the data and is closely oriented to the results of the previous step, the evaluation.

In this step, the limitations of the data are addressed in order to improve the quality of the data and make it usable for the research community, while ensuring that it provides reliable results.

Corpora

Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2023-24

Cristina Bosco

Types of Corpora

Corpora can be classified according to several dimensions.

Language: the most of corpora are **monolingual** but an increasing amount of **bilingual or multilingual** corpora is available.

Content and organisation: bilingual and multilingual corpora can be **parallel** or **aligned**.

The notion of **comparability** can be referred to all types of corpora.

Comparable corpora

Two **comparable** corpora contain components in the same or in different languages that have been collected using the **same sampling method**, such as the *same proportions* of the texts of the *same genres* in the *same domains* in the *same sampling period*.

Comparable corpora are texts originally produced (not translated) in the respective languages which consist of independent texts which are therefore "similar".

Their comparability lies in the similarity of their sampling frames.

Parallel corpora

Two **parallel** corpora contain native language (L1) **source texts** and their (L2) **translations**.

The sampling frame is automatically the same for all the languages in the corpora.

Among the most important initiatives for the collection and distribution of parallel corpora there is OPUS (https://opus.nlpl.eu/), a growing collection of translated texts from the web.

OPUS texts are free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus.

Aligned corpora

Two **aligned** corpora are parallel corpora in which source texts and their translations are associate, annotating the correspondences between the two at the sentence or word level.

The alignment is a time-consuming task that can be manually or automatically applied.

The automatic alignment of parallel corpora is possible only for some language pairs, but it can still be a challenge for others.

Using corpora

Aligned corpora are particularly useful for Machine Translation. Modern approaches to Machine Translation are in fact based on statistical methods, and an aligned corpus is the best source to extract for each word the probability that a word W-1 will be translated with W-2 and not with W-3.

If no aligned corpus is available, a **parallel** corpus can be used for the same purpose.

Using corpora

Comparable corpora are useful in several NLP tasks.

For instance, if you train a model on a corpus X, you can expect that the model works almost as well on data from a comparable corpus as it does for X.

Bilingual knowledge to be included in word embedding must be extracted from bilingual corpora.
Especially accurate information can be extracted from **aligned corpora**, **parallel corpora**, or **comparable**, with a variable degree of reliability.