Consorzia, campaigns and resources

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

Cristina Bosco

Overview

- Distribution of resources: consortia, associations and copyrights

- Evaluation campaigns: general information and participation

Distribution of resources

Resources are published and distributed in repositories from which they can be downloaded.

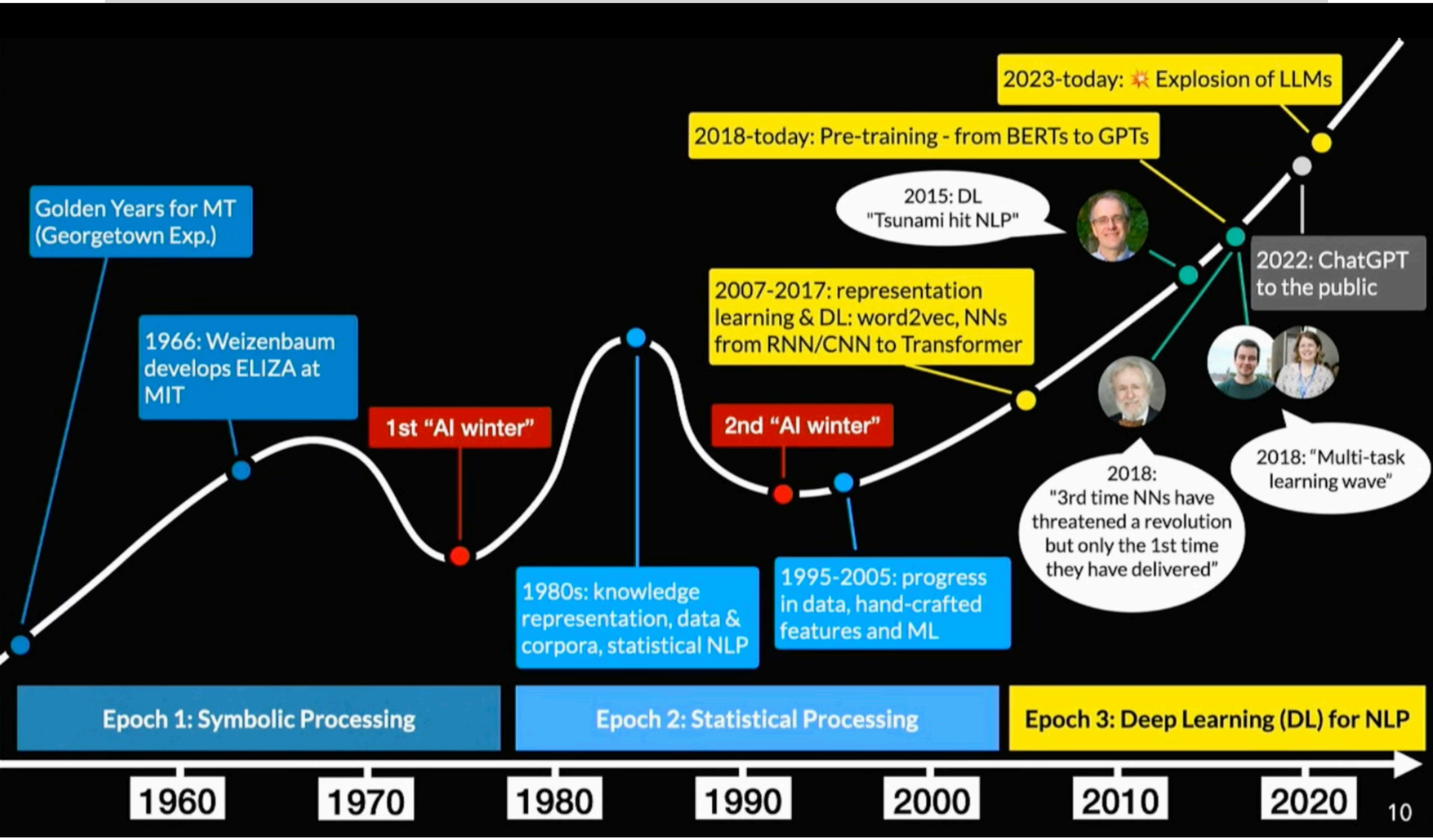
This improves the possibility to access them for researchers that want study them or do experiments using their data.

A bit of history

Thanks to the emergence of systems based on machine learning and statistical methods, the first winter of language technologies and NLP, defined in the US by the ALPAC report and the widespread criticism of MT, has come to an end.

What followed was a resurgence of interest and an insatiable hunger for language resources. Several research groups all around the world started working on the development of novel resources

A bit of history



A bit of history

By the early 1990s, as a result of DARPA's (Defense Advanced Research Projects Agency*) successes, language technology research was described optimistically as having "*useful present-day systems and realistic expectations of progress*".

*DARPA is the agency of the United States Department of Defense responsible for the development of emerging technologies for use by the military.

A bit of history

But **language resources soon became a bottleneck**: the extent to which language resources, or their absence or limitations, might affect the progress of language technologies became a major concern.

The development of language resources was particularly costly at this time, as all resources had to be developed from scratch, by designing all their features and then creating through mostly manual labor.

A bit of history

Not even the largest companies could easily afford enough of the needed data to satisfy their research and development needs.

Research institutions, smaller companies and universities were at risk of being frozen out of the language technology development almost entirely.

In Europe also there were a growing awareness of the need for publicly available very large common corpora.

LDC

In the USA, with DARPA support, the **Linguistic Data Consortium** was founded in 1992 at Penn University

https://www.ldc.upenn.edu/

The objective of this initiative was to provide a mechanism for large-scale **development** and widespread **sharing of resources** for research in language technologies.

Today on the LDC Catalog platform resources are available for addressing a variety of tasks.

LDC

§ handwriting recognition
§ entity, event, relation extraction & coreference
§ information retrieval
§ knowledge base population
§ language identification
§ language modeling
§ machine translation
§ parsing, POS tagging & other NLP
§ pronunciation modeling
§ question-answering
§ semantic role labelling
§ sentiment detection
§ speaker diarization, identification
§ speech activity detection
§ speech recognition
§ summarization

LDC

Since its foundation the consortium releases resources to a large number of subscribers who pay an annual subscription of thousands dollars.
One of the earlier and most important resources is the Penn Treebank.

Today's LDC Catalog includes more than 900 corpora in 107 linguistic varieties, including recent additions also in less resourced languages Dari, Georgian, Icelandic, Kazakh, Kurdish, Nahuatl, Persian, Pushto, Russian, Turkish Ukrainian, Uzbek and Zulu.

ELRA

ELRA, the European Language Resources Association
http://www.elra.info/en/

has been founded in 1995 and is a non-profit organisation whose main mission is to **make** language resources for human language technologies **available** to the community at large.

ELRA organizes every two year the **Language Resources and Evaluation Conference** (LREC), which is the most important event for the research community that works on resources.

Among the initiatives of ELRA, it is especially interesting the LRE journal: https://www.elra.info/dissemination/language-resources-and-evaluation-journal/

ELG

The European Language Grid
(https://live.european-language-grid.eu/grid)

is a platform that **offers access** to a multitude of assets related to language technology, including commercial and non-commercial cloud services for all European languages, data resources such as models, datasets, lexica, terminologies or grammars, but also information on projects, events and associations.

Sharing linguistic resources

Where to find a resource?

Resources may be distributed by consortia or associations, such as LDC or ELRA, or directly by the authors (in GitHub repositories or websites), or within evaluation campaigns.

The access to a particular resource is often provided by more than one platform or website.

For example, on SketchEngine you can find resources a large variety of corpora (https://www.sketchengine.eu/).

Sharing linguistic resources

Do you have to pay to use a language resource?

Many language resources are shared **free of charge**, while other resources are distributed **for a fee**.

Academic research groups usually share the resources they develop for free, while consortia, such as LDC or ELRA, didn't.

The advantage of distribution made by an institution responsible for this purpose, such as consortia, is the maintenance of the resources: they are carefully maintained and remain accessible also for long time, while sometimes resources shared by research groups are not.

Sharing linguistic resources

What must be done to use a language resource?

All resources are released under a licence:

- for **free** resources, it may be a licence such as Creative Commons; this licence defines the possible uses of the resource (usually non-commercial) and is issued by the copyright holder to allow anyone to use the resource in any way that complies with that licence, carefully **citing the author and the licence**

- for the other resource, an **agreement** must be signed and a **fee** paid to the author or distributor.

Creative Commons

"Creative Commons is an international nonprofit organization that empowers people to grow and sustain the thriving commons of shared knowledge and culture we need to address the world's most pressing challenges and create a brighter future for all."

The organization releases several copyright licenses known as Creative Commons licenses, free of charge to the public.
The authors of creative works using them can communicate which rights they reserve for themselves and which rights they renounce for the benefit of other people.

The list and description of the licenses is at:
https://creativecommons.org/share-your-work/cclicenses/

Ethical issues

Sharing linguistic data also means carefully aligning them with privacy policies, such as the GDPR.

The General Data Protection Regulation (GDPR) is the European law on data privacy and security.
It is the toughest privacy and security law in the world. Although it was drafted and adopted by the European Union, it imposes obligations on organizations everywhere as long as they target or collect data related to people in the EU.
This regulation went into effect on May 25, 2018.

Ethical issues

For example, to be compliant with the GDPR, you must not publish data that contains sensitive information about an individual.

This means that you must be very careful when collecting data that is actually made public by the author (e.g., posts on social media) and then **pseudonymize** it (replacing "*John Smith*" with "*name-of-person*") to avoid incidentally identifying the author.

Evaluation campaigns

An evaluation campaign is the place in which linguistic resources are shared and evaluation exercises are organised using them.

The goal of campaign is to compare results,
to test the validity of resources and models
and to share language technologies with the research community and industry both.

Evaluation campaigns

Evaluation is applied in evaluation campaigns, in which resources and systems are formally tested.
An evaluation campaign is organized into several different **shared tasks** and the organisation of each task comprises in turn the following phases:

1) **training phase**: the organisers of the task distribute a large set of data (***training set***) to the participants. Only the training set must be used to train systems in *closed* tasks, while also other resources (selected by each participant) can be used in the *open* ones.

Evaluation campaigns

The ***training set*** includes the samples that a system must use to build a model of the data to be later processed (those included in the test set).

But this set is also used for evaluating the validity of the model during its development.

When a portion of the training set is taken apart and used for testing the model during its development, before the final evaluation based on the test set, it is called ***development set***.

Evaluation campaigns

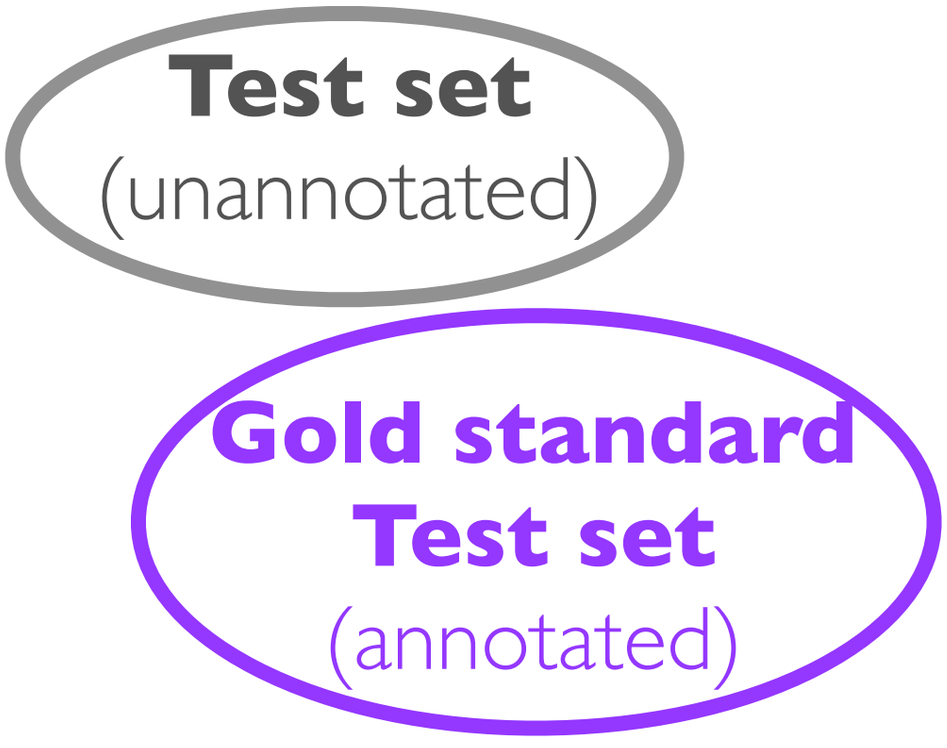
The development set is usually composed by 20-25% around of the data of the training set.

The **training set** is released by the organisers in an annotated version, which can be gold (if the annotation has been manually checked) or silver (if the annotation has been only automatic).

The **development set** is used in the unannotated version for testing the model, and in the annotated version to evaluate model results.

Evaluation campaigns

The **test set** is usually composed by 20-25% around of the data of the training set. To be used in the official evaluation of participant results, the test set is usually a gold standard, carefully annotated and manually checked.
The test set is firstly released by the organisers in an unannotated version, to be used in the official test of participant systems.



The test set is firstly released by the organisers in an **unannotated** version, to be used in the official test of participant systems.

The gold standard test set, **annotated**, is used by the organisers in the evaluation and then made public.

Evaluation campaigns

2) **testing phase**: an unannotated version of the gold standard ***test set*** is distributed to the participants that have to run their systems on it and to provide their results to the organisers; usually more than one run can be provided by each participant

3) **evaluation phase**: the results generated by participant systems on the test set are compared with those in the gold standard test set to calculate the given evaluation metrics for each participant system. The gold standard test set is made public.

Evaluation campaigns

As an **organizer** you must:

- create the (gold standard) **training set**, the **guidelines** and the **gold standard test set**

- share with participants the training set (organised or not in training and development) and the **evaluation scripts**

As a **participant** you must:

- train your system on the training set, test it on the development set, evaluate results; tune your system, test it on the development set, evaluate results … until you achieve good metric scores

Evaluation campaigns

• share with participants the unannotated test set	• run your system on the test set (usually more than one run is accepted, each run is the result of a different tuning of your model)
• evaluate participant results against gold standard test set • publish the official results	• share your results (one or more run) with the organisers
• write a report of the task	• write your participation report
• organize the final workshop	• participate to the final workshop

Evaluation campaigns

Some evaluation campaigns is an effort of a national computational linguistics associations, and is focused on a specific language, such as:

ITALIAN >
https://www.evalita.it/

GERMAN >
https://germeval.github.io/

SPANISH >
https://sites.google.com/view/iberlef2022/home?
(the name is IberEval until 2018 now is IberLEF)

Evaluation campaigns

Some evaluation campaigns is organised about some (set of) specific type of tasks by international associations, such as

SemEval (Semantic Evaluation - semantics) > https://semeval.github.io/

CoNLL (Conference on Computational Natural Language Learning - morphology and syntax) > https://www.conll.org/previous-editions