Traceability, Reproducibility, Replicability

What It Means for Computational Linguistics

Nancy Ide

Department of Computer Science

Vassar College

Poughkeepsie, New York USA

Background

Motivation

Failures of reproducibility have been a concern in the global scientific community

- Wikipedia page on the topic entitled "Replication Crisis"
  - Description of some of the most worrying results and links to the relevant studies
- Survey conducted by *Nature* in 2016
  - More than half of over 1,500 participating scientists claim that there is a "significant reproducibility crisis."

Motivation

Why is reproducibility necessary?

- Need to critically assess
  - correctness of scientific claims
  - conclusions drawn by other scientists

But there is ample evidence that it is rarely possible to adequately validate studies described in academic publications

Terminology

- Given the importance of validating and verifying work described in scholarly publications, expect that there would be at least a broad consensus about the terminology used

- However, even a cursory review of the literature shows that no such consensus exists
  - “At least the following *three* terms used frequently to refer to the same *two* concerns: **replicability**, **repeatability**, and **reproducibility**” (Cohen et al., 2018)

Brief History

- Historically, scientists expected that
  - Experiments are described in sufficient detail that others can **follow the steps** and **obtain the same results** within the margins of experimental error
  - Insights into nature (e.g., measurement of the speed of light, propagation of action potentials along axons) could be independently confirmed using **different experimental means**
- Doubts about the interpretation of certain results gave rise to new branches of science (e.g. Schrödinger, 1915)
  - Experimental scientists developed a systematic approach over decades, well-established in the literature and international standards

Hans E. Plesser , Reproducibility vs. Replicability: A Brief History of a Confused Terminology, Frontiers in Neuroinformatics 11, 2017

Age of Computers

- Attention to experimental error took back stage when scientists begin to use digital computers to perform simulations and data analysis
  - **Assumed results obtained were exact** and could be trusted if algorithms and methods were suitable to the problem
- Little attention paid to
  - Correctness of implementation
  - Potential for error
  - Variation introduced by system soft- and hardware
  - Difficulties to reconstruct how an experiment had been performed

Early Attempts to Address the Problem

- Claerbout and Karrenbach, 1992
  - Defined **reproducing** to mean “running the same software on the same input data and **obtaining the same results**”

> **“[j]udgement of the reproducibility of computationally oriented research no longer requires an expert—a clerk can do it”**

  - Defined **replicating** to mean “writing and then running new software based on the description of a computational model or method provided in the original publication, and **obtaining results that are similar enough** …”
- Followed up by Donoho et al., 2009 and Peng, 2011

Claerbout/Donoho/Peng (CDP) convention

- **Reproducible research**
  - Authors provide all the necessary data and the computer codes to run the analysis again, **re-creating the results**
- **Replication**
  - A study that arrives at the same scientific findings as another study, **collecting new data** (possibly with different methods) and **completing new analyses**

Unfortunately…

This terminology at odds with terminology long established in experimental sciences

CDP: reproducibility

> …modern convention makes a careful distinction between **reproducibility** and **repeatability**. …student A …would do the five replicate titrations in rapid succession …. The same set of solutions and the same glassware would be used throughout, the same temperature, humidity and other laboratory conditions would remain much the same. In such circumstances, the precision measured would be the **within-run precision**: this is called the **repeatability**. Suppose, however, that for some reason the titrations were performed by different staff on five different occasions in different laboratories, using different pieces of glassware and different batches of indicator …. This set of data would reflect the **between-run precision of the method**, i.e. its **reproducibility**.
> (Miller and Miller, 2000)

CDP: replicability

Note: here reproducibility refers to errors arising in different laboratories and equipment but using the same method, more restricted definition than used elsewhere

Giving rise to….

The Terminology Wars

Heated Debate

WHICH TERMINOLOGY IS THE PROPER ONE?

DISCUSSION ON “R-WORDS” ON GITHUB (ROUGIER ET AL., 2016)

CONFUSION MAY ARISE FROM DRUMMOND’S (2009) SWAPPING OF DEFINITIONS

Chris Drummond, 2009: "Replicability is not Reproducibility: Nor is it Good Science",

- Attempted to bring terminology in computational science **in line with the experimental sciences**
- But also argued that *one should not focus on collecting computer-experimental artifacts to ensure that simulations and analyses can be re-run*

> “I want to … [separate] the notion of **reproducibility**, a generally desirable property, from **replicability**, its poor cousin. I claim there are important differences between the two. **Reproducibility** requires changes; **replicability** avoids them. Although **reproducibility** is desirable, I contend that the impoverished version, **replicability**, is one not worth having.”

CDP: replicability

CDP: reproducibility

It Becomes More Complicated…

After this, not uncommon to see *reproducibility* and *replicability* used interchangeably in the same paper

- “This experience motivated the creation of a way to encapsulate all aspects of our in silico analyses (3) in a manner that would facilitate independent **replication** by another scientist (4). Computer and computational scientists refer to this goal as “**reproducible** research” … (Mesirov, 2010)
- “**Reproducibility**, or **replicability**, is the quality of a scientific experiment that can be performed independently several times and yield the exact same results on each iteration.” (Névéol and Grouin, 2016)

Victoria Stodden et al., Eds., 2014
Implementing Reproducible Research

The first comprehensive review of the field in book form

> "**Replication**, the practice of independently implementing scientific experiments to validate specific findings, is the cornerstone of discovering scientific truth. Related to replication is **reproducibility**, which is the calculation of quantitative scientific results by an independent scientist using the original datasets and methods. **Reproducibility** can be thought of as a different standard of validity from **replication** because it foregoes independent data collection and uses the methods and data collected by the original investigator." (Preface, p. vii)

In the Claerbout/Donoho/Peng camp

However…

Some authors of chapters in the book don’t make a clear distinction between the terms ‘reproduce/replicate,’ and often use ‘reproducibility‘ as an umbrella term

- E.g., chapter 11 (by the Open Science Collaboration): “. . . narrowly, **reproducibility** is the *repetition* of a simulation or data analysis of existing data by re-executing a program. More broadly, **reproducibility** refers to *direct replication*, an attempt to replicate the original observation using the same methods of a previous investigation but collecting new [data].”

✓ U.S. National Science Foundation

Bollen K, Cacioppo JT, Kaplan RM, et al. 2015

- **Reproducibility**: The ability to duplicate the results of a prior study using the **same materials and procedures** as were used by the original investigator
- **Replicability**: The ability to duplicate the results of a prior study if the **same procedures** are followed but **new data** are collected
- **Generalizability**: Whether the results of a study apply in other contexts or populations that differ from the original one (also referred to as **translatability**)

In the Claerbout/Donoho/Peng camp

Association of Computing Machinery (ACM) Result and Artifact Review and Badging (2016)

A badging system for articles complying with various standards of code and data sharing

- **Repeatability**
  - **Same team, same experimental setup**
    - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials.
- **Replicability** ← **CDP/NSF: reproducibility**
  - **Different team, same experimental setup**
    - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials.
- **Reproducibility** ← **CDP/NSF: replicability**
  - **Different team, different experimental setup**
    - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials.

A New Lexicon

Goodman et al. (2016)

- "…basic terms—reproducibility, replicability, reliability, robustness, and generalizability—are not standardized."
- To solve the terminology confusion, proposes a new *lexicon for research reproducibility*:

  CDP/NSF: reproducibility ✓

  - **Methods reproducibility**: ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results

  CDP/NSF: replicability

  - **Results reproducibility**: production of corroborating results in a new study, having followed the same experimental methods
  - **Inferential reproducibility**: **draw the same conclusions** from either an independent *replication* of a study or a reanalysis of the original study

Comparison

Experimental science/ Drummond	Goodman	Claerbout/ Donoho/Peng	ACM	NSF
			Repeatability	
Replicability	Methods reproducibility	Reproducibility	Replicability	Reproducibility
Reproducibility	Results reproducibility	Replicability	Reproducibility	Replicability
	Inferential reproducibility			

Terminologies for Reproducible Research Lorena A. Barba, 2018

- Classifies terminology use via decision tree:
  - **A**—make no distinction between the words *reproduce* and *replicate*
  - **B**—use them distinctly
    - **B1**: *reproduce =* same data+same methods=same results
    - **B2**: *replicate =* same data+same methods=same results

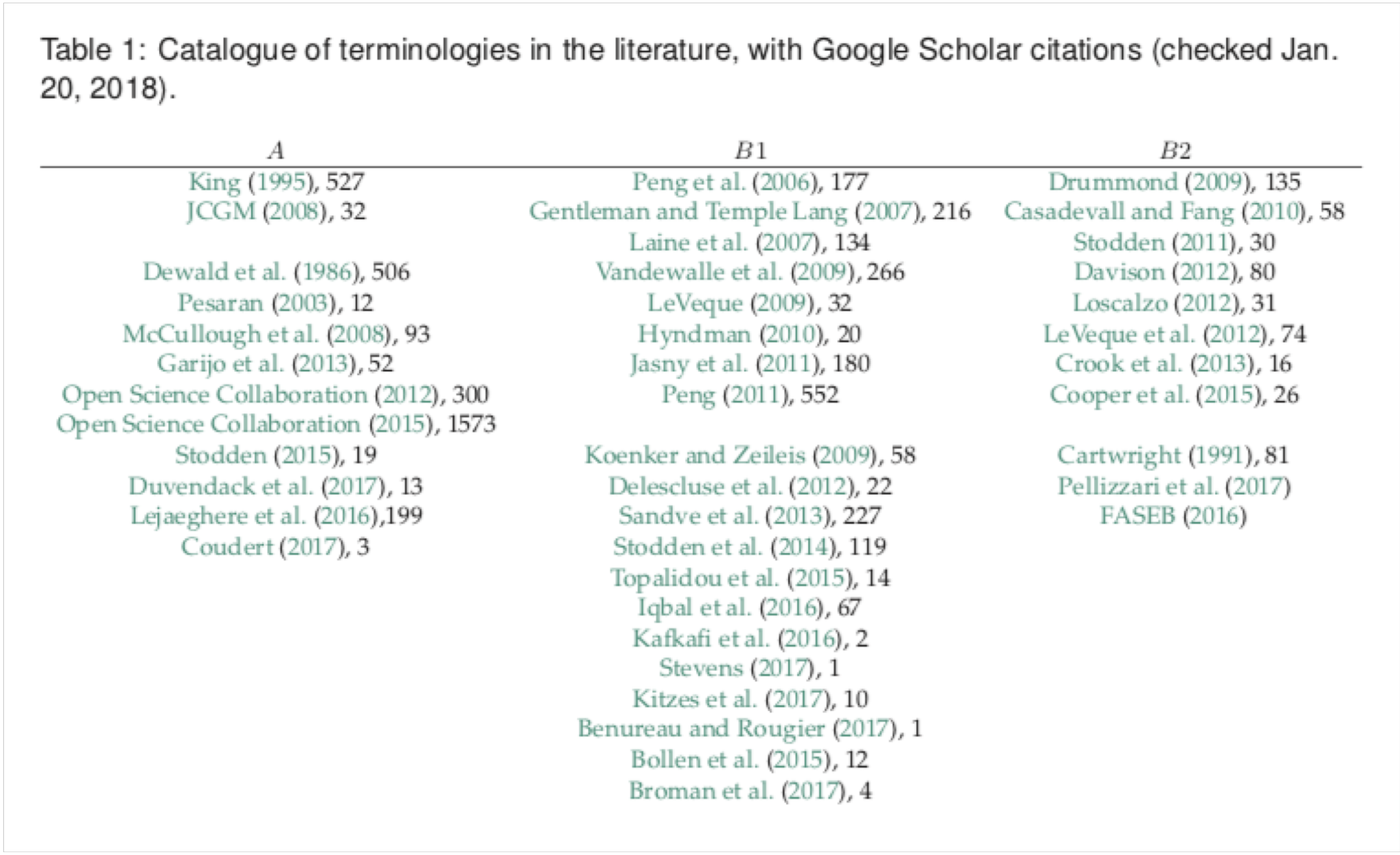


Table 1: Catalogue of terminologies in the literature, with Google Scholar citations (checked Jan. 20, 2018).

A	$B1$	$B2$
King (1995), 527	Peng et al. (2006), 177	Drummond (2009), 135
JCGM (2008), 32	Gentleman and Temple Lang (2007), 216	Casadevall and Fang (2010), 58
	Laine et al. (2007), 134	Stodden (2011), 30
Dewald et al. (1986), 506	Vandewalle et al. (2009), 266	Davison (2012), 80
Pesaran (2003), 12	LeVeque (2009), 32	Loscalzo (2012), 31
McCullough et al. (2008), 93	Hyndman (2010), 20	LeVeque et al. (2012), 74
Garijo et al. (2013), 52	Jasny et al. (2011), 180	Crook et al. (2013), 16
Open Science Collaboration (2012), 300	Peng (2011), 552	Cooper et al. (2015), 26
Open Science Collaboration (2015), 1573		
Stodden (2015), 19	Koenker and Zeileis (2009), 58	Cartwright (1991), 81
Duvendack et al. (2017), 13	Delescluse et al. (2012), 22	Pellizzari et al. (2017)
Lejaeghere et al. (2016),199	Sandve et al. (2013), 227	FASEB (2016)
Coudert (2017), 3	Stodden et al. (2014), 119	
	Topalidou et al. (2015), 14	
	Iqbal et al. (2016), 67	
	Kafkafi et al. (2016), 2	
	Stevens (2017), 1	
	Kitzes et al. (2017), 10	
	Benureau and Rougier (2017), 1	
	Bollen et al. (2015), 12	
	Broman et al. (2017), 4	

Terminology by Discipline

Detailed information?

Stodden (2014)

- **Computational reproducibility**
  - Detailed information is provided about code, software, hardware and implementation details
- **Empirical reproducibility**
  - Detailed information is provided about non-computational empirical scientific experiments and observations
    - Enabled by making data freely available together with details of how the data was collected
- **Statistical reproducibility**
  - Detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc.

Terminology by Discipline

Economics, Political Science

CDP/NSF: reproducibility

Hamermesh (2007)

- **Statistical replication**: re-analyzing the same data with the same model and estimation parameters
- **Scientific replication**: use "different sample, different population, and perhaps **similar but not identical model** [. . .] and, indeed, comprises most of what economists view as replication."

Beyond CDP/NSF: replicability

Terminology by Discipline

Neuroscience

Crook et al. (2013)

- **Internal replicability**: the original authors or someone else in the same group can re-create the results, re-executing the same software
- **External replicability**: a reader of published results can re-create them using the data and code supplied by the authors

CDP/NSF: reproducibility

- **Cross replicability**: running the same model but with **different software**

No mention of data

- **Reproducibility**: “the boundary line between cross-replicability and reproducibility is not always clear.”

CDP/NSF: replicability

Terminology by Discipline

Psychology, Experimental Biology

✓ ACM

The American Psychological Association (Appelbaum et al., 2018)

- ‘**Replication** studies’ : complete studies (including data collection) meant to confirm the findings of another

The Federation of American Societies for Experimental Biology ✓

- **Replicability**: the ability to **duplicate (i.e., repeat) a prior result** using the **same source materials and methodologies**. This term should only be used when referring to repeating the results of a specific experiment rather than an entire study;
- **Reproducibility**: the ability to **achieve similar or nearly identical results** using **comparable materials and methodologies**. This term may be used when specific findings from a study are obtained by an independent group of researchers.

Terminology by Discipline

Statistics

✓ **CDP/NSF**

The American Statistical Association (Broman et al., 2017)

- Recommendations for funding agencies on supporting reproducible research
  - **Reproducibility**
    - A study is reproducible if you can take the **original data and the computer code** used to analyze the data and **reproduce all of the numerical findings** from the study.
  - **Replicability**
    - This is the act of **repeating an entire study**, independently of the original investigator **without the use of original data** (but generally **using the same methods**).

Terminology by Discipline

Computational research

✓ **CDP/NSF**

- ReScience (Rougier et al., 2017)
  - “a peer-reviewed journal that targets computational research and encourages the explicit replication of already published research.”
  - **Reproducing** the result of a computation means running **the same software** on the **same input data** and obtaining the **same results**. The goal of a reproduction attempt is to verify that the computational protocol leading to the results has been recorded correctly.
  - **Replicating** a published result means **writing and then running new software based on the description of a computational model or method** provided in the original publication, and obtaining **results that are similar enough to be considered equivalent**.

Terminology by Discipline

(based on Barba, 2018)

No distinction between *reproduce* and *replicate*	***Reproduce* = same data+same methods=same results**	***Replicate* = same data+same methods=same results**
Political science	Signal processing	Microbiology, immunology
Economics	Scientific computing	Computer Science
	Econometry	Psychology
	Epidemiology	Experimental Biology
	Clinical studies	
	Internal medicine	
	Physiology (Neuro)	
	Computational biology	
	Biomedical research	
	Statistics	

Reproducibility as Open Code and Data

Gentleman and Temple Lang (2007)

“By **reproducible** research, we mean research papers with accompanying software tools that allow the reader to directly **reproduce** the results and employ the computational methods that are presented in the research paper.”

Vandewalle et al. (2009)

“A research work is called **reproducible** if all information relevant to the work, including, but not limited to, text, data and code, is made available, such that an independent researcher can reproduce the results.”

LeVeque (2009)

“The idea of ‘**reproducible** research’ in scientific computing is to archive and make publicly available all the codes used to create a paper’s figures or tables, preferably in such a manner that readers can download the codes and run them to **reproduce** the results.”

Reproducibility as Open Code and Data

Donoho et al. (2009)

Define reproducible computational research as that “in which all details of computations—code and data—are made conveniently available to others.”

For NLP, Pedersen (2008):

“…releasing software that makes it easy to reproduce and modify experiments should be an essential part of the publication process, to the point where we might one day only accept for publication articles that are accompanied by working software that allows for immediate and reliable reproduction of results.”

Reproducibility as Open Code and Data

Stodden et al. (2013) place **computational reproducibility** on a spectrum with five categories that vary by degree of availability of code and data:

Reviewable Research

Replicable Research

Confirmable Research

Auditable Research

Open or Reproducible Research

Reproducibility
Openness

Repeatability

- Taylor and Kuyatt, 1994
  - **Repeatability** is the precision over successive measurements of the same quantity, with **everything kept the same (even the operator)**, over a short period of time. **Reproducibility** of measurements involves **changing at least one condition**, e.g., the instrument, the location, or the operator to measure the *same* physical quantity.

ACM: repeatability

~ CDP/NSF: replicability

✓

- Dalle, 2012
  - Conditions for the **repeatability** of results measurements require the same observer, the same instrument used in the same conditions and the same location

ACM: repeatability

- Most relevant for fields such as Chemistry, etc.?

Repeatability

Collberg et al., 2014

- **Reproducibility**: the independent confirmation of a scientific hypothesis through reproduction by an independent researcher/lab

  “In the context of computer science research, reproducibility can usefully be replaced by the concept of repeatability.”

- Three types of “weak repeatability”
  - *Highest level:* ability of a system to be acquired and then built in 30 minutes or fewer
  - *Next level:* ability of a system to be acquired, and then built, regardless of the time required to do so
  - *Lowest level:* ability of a system to be acquired, and then either built, regardless of the time required to do so, or the original author’s insistence that the code would build, if only enough of an effort were made.

✓

Reproducibility in Computational Linguistics

Does the reproducibility crisis extend to natural language processing?

Despite appearances, numerous applications reported in the literature turn out to be uncompilable, unusable, unobtainable, or otherwise not reproducible

Pedersen (2008) “Empiricism Is Not a Matter of Faith”

Often impossible to obtain the relevant data and software in order to reproduce a study

The Usual Suspects

- Failures of reproducibility in the field are due to many of the same issues as for other disciplines
- Lack of
  - Sufficient detail about methods
  - Access to or ability to use the exact data used
  - Access to or ability to use the exact software used

Terminology

✓ **ACM**

- Fokkens et al. (2013)
  - **Reproducibility**: Ability to reproduce the **same answer** to a research question by **different means**, perhaps by re-implementing an algorithm or evaluating it on a new (in domain) data set.
  - **Replication** : simply involves running the **exact same system** under the **same conditions** in order to get the exact **same results** as output.
- Wieling et al. (2018)
  - “[W]ith **reproduction** (or **reproducibility**), we denote the **exact re-creation of the results** reported in a publication using the **same data and methods**.”

Terminology

An obeservation:

- Cohen et al. (2018)
  - "**Replicability** or **repeatability** is a property of an *experiment*: the ability to repeat—or not—the experiment described in a study."
  - "**Reproducibility** is a property of the *outcomes* of an experiment: arriving—or not—at the same conclusions, findings, or values."

Hierarchy of Reproducibility

Cassidy et al., 2017

1. All computations clearly described in terms of well-known methods
2. Specific software packages are referenced
3. Details of which functions within packages were used
4. Exact settings used to run the computation are given
5. A copy of the scripts used for data processing is available
6. Software can be downloaded and executed
7. Results are the same as those in the paper

The farther down this list one can go, the stronger the claim that the research is reproducible

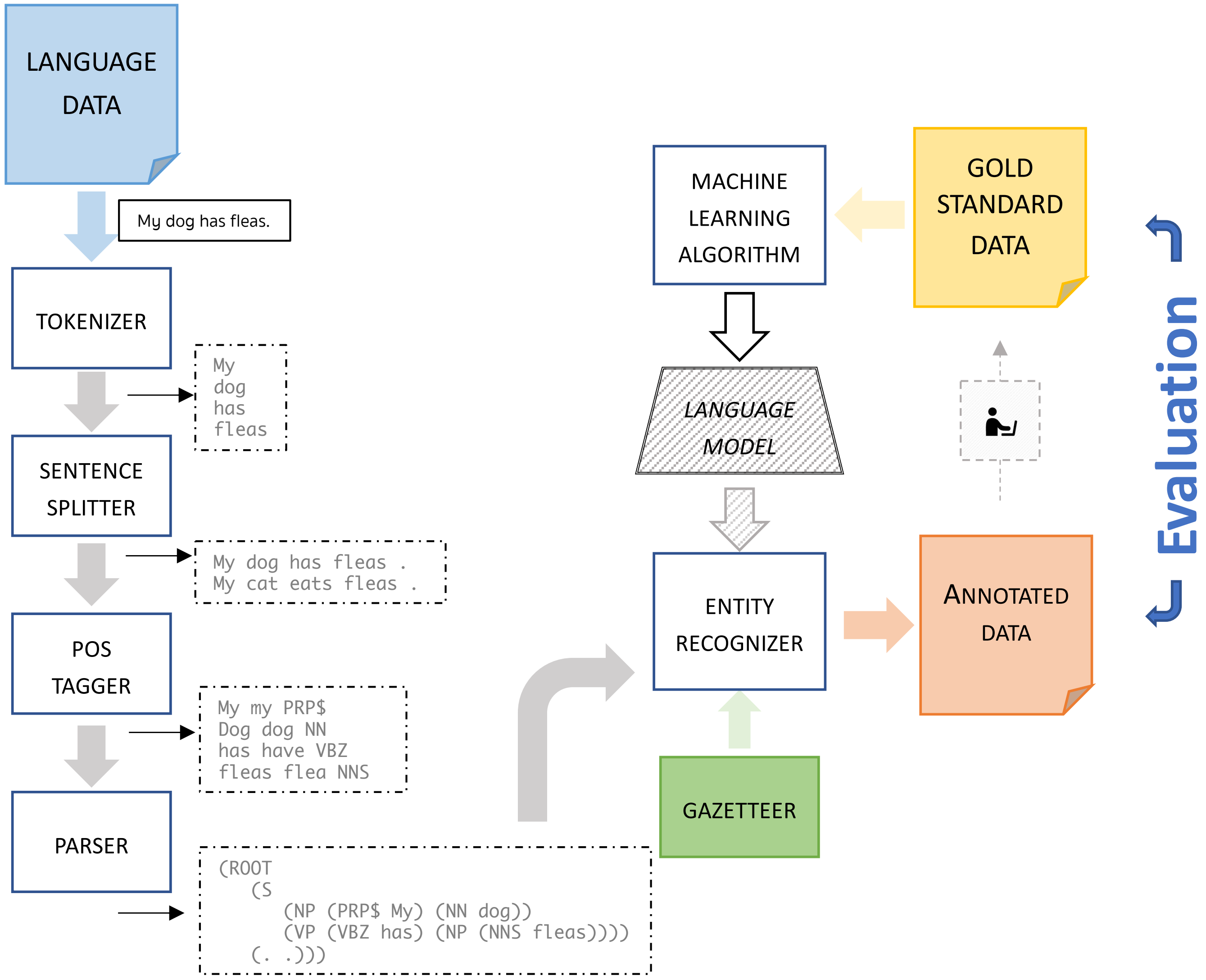
Computational Linguistics Research

- In most cases, goal is to model or explore corpora of **language data**
  - Few papers in the field are without reference to a digital collection of language data, whether small or large
  - Collecting and preparing data is often the most time-consuming part of a research project

Computational Linguistics Research

- Relies on corpora, but also supporting **language resources** such as lexicons, dictionaries, ontologies, and the like
- Typical experiment generates some result for an enabling technology (software for identification of some linguistic phenomenon, e.g., named entity recognition, text entailment, coreference resolution…) or application (information extraction, machine translation…)
- Results consist of statistics such as **precision, recall, f-score**
- Often report an improvement of 2-3% over state-of-the-art

Example flow in CL studies



```
My dog has fleas.
```

```
My
dog
has
fleas
```

```
My dog has fleas .
My cat eats fleas .
```

```
My my PRP$
Dog dog NN
has have VBZ
fleas flea NNS
```

```
(ROOT
   (S
      (NP (PRP$ My) (NN dog))
      (VP (VBZ has) (NP (NNS fleas))))
   (. .)))
```

Issues for Reproducibility

- Experiments often consist of **cascade of processing steps** and configurations
  - Many steps (preprocessing techniques, alignment parameters, translation rule extraction parameters, language model parameters, list of features used) invariably omitted in publications
- Results may be obtained by a **single system built to perform a given task** (e.g., Entity Linking, Sentiment Analysis)
  - Often involve **custom code** (may build on widely available components such as the Stanford NLP Tools (Manning et al., 2014) and freely available architectures (e.g., GATE, UIMA)
- Workflow may make use of a number of software components, **combine manual and automatic steps** towards an end result

Data Availability is a Critical Factor

- Many widely shared data-sets used as the raw material for research are available
  - Through services like LDC and ELRA (some for a fee)
  - Datasets from activities such as shared tasks, which prepare manually annotated corpora for testing and evaluation, usually available from various websites
- However
  - Many datasets are only available via download from a research group website, which may become unavailable at any time
  - Many datasets are restricted by copyright or other considerations
  - For sensitive data in the medical, intelligence, and law enforcement domains, the problem of unavailability of data will probably never be addressed in such a way as to facilitate reproducibility

Wikipedia and the “Web as Corpus”

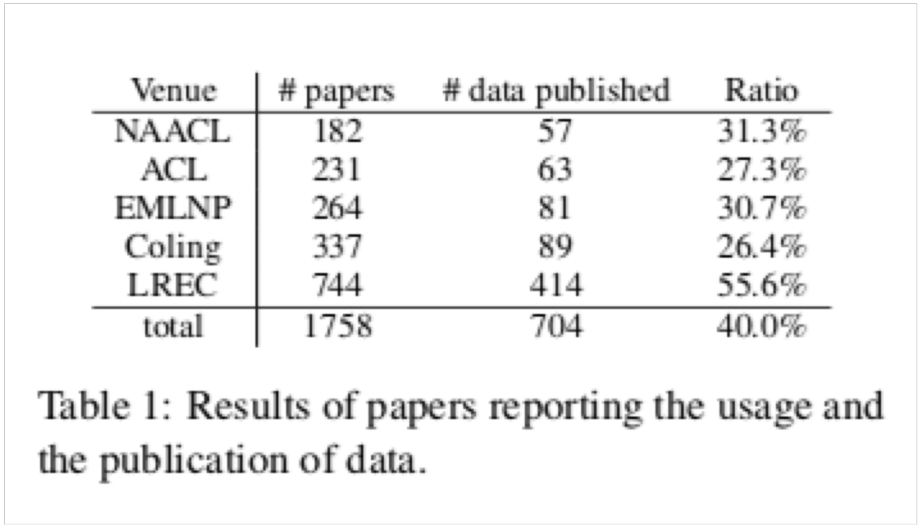
- Because large datasets of language data are difficult to manually collect, there is **widespread use of Wikipedia** and other web-crawled datasets as a source
- Wikipedia is a **constantly changing resource** and even snapshots taken days apart can vary significantly[*]
  - General references to Wikipedia as a source does not mean that we have access to exactly the data used in the study
    - Studies rarely give a time that the snapshot used in the study was taken
  - In some cases, selection has been done on the data to remove outliers or noisy data
    - Exact input for the experiment is therefore not clear
- **Web-crawled data is not necessarily stable** and therefore may not be reproducible, often requires **a lot of (variable) pre-processing** to use

[*] To partially address this, a number of published snapshots of Wikipedia have been made available specifically as NLP corpora (e.g., Reese et al., 2010)

Mieskes (2017)

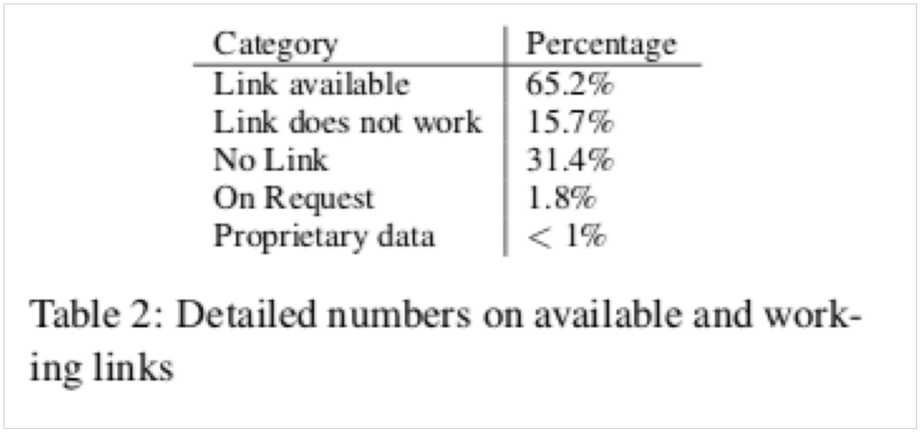
- Quantitative analysis of publications in the NLP domain on collecting, publishing, and **availability of research data**
- Investigated how often studies published at various computational linguistics conferences provided a link to the data
- Findings
  - 40% of the papers collected new data or changed existing data
  - Only in about 65% of these papers was a link to the data provided
    - 18% of links did not work
  - A wide range of publications rely on data crawled from the web
  - Few give details on how potentially sensitive data was treated

Mieskes (2017) Summary of Results



Venue	# papers	# data published	Ratio
NAACL	182	57	31.3%
ACL	231	63	27.3%
EMLNP	264	81	30.7%
Coling	337	89	26.4%
LREC	744	414	55.6%
total	1758	704	40.0%

Table 1: Results of papers reporting the usage and the publication of data.



Category	Percentage
Link available	65.2%
Link does not work	15.7%
No Link	31.4%
On Request	1.8%
Proprietary data	$<$ 1%

Table 2: Detailed numbers on available and working links

Wieling, et al. 2018

Focus on **availability of source code**

- Change from 2011 to 2016:
  - Percentage of papers providing a (working) link to the source code approximately doubles (18.6% to 36.2%)
  - Requesting source code (if not already provided) unlikely to be successful
    - About a third of the requests was (or could be) granted
- Five studies from each of 2011 and 2016
  - At most 60% reproducible when not enforcing an exact reproduction
  - Exact reproduction only for a single study (from 2011)

Language Data

Not your average scientists’ data

Gathered from a myriad of sources

Much more susceptible to changes, normalization

Must deal with character sets, fonts, etc., conversion from formats like html, pdf

Typically undergoes substantial processing to identify low-level linguistic elements before getting to the main tasks

Data Preparation

(The Biggest Reproducibility Problem?)

- Choice of which steps we perform, and how each of these steps is carried out exactly are part of our experimental setup
- But the **pre-processing steps are rarely outlined or documented**
  - Various kinds of normalization, elimination or modification of special characters, lowercase the whole corpus, correction of spelling errors, typos…
  - Anonymization of sensitive data
  - Annotated corpora present additional problems: annotation bugs are fixed, formats or codes changed

Domain-specific problems

- Natural language processing research is **primarily published in conference proceedings**, not journals
  - Conference papers routinely have **page limits**: typically not enough space to capture all details, and they are generally not the core of the research described.
- Little or **no tradition in the community of publishing reproduction attempts**—bias strongly in favor of novel methods
  - Question on many review forms: ‘how novel is the presented approach?’
- **Not enough (academic) credit** is gained from making resources available
- **Wide range of technical expertise** of researchers in this field
  - Problem of accessibility for tools that might be impossible to run without deep technical knowledge
  - Impedes the application of tools in a cross-disciplinary manner

Reproducibility Experiments in CL

Bikel (2004) attempted to reproduce the parsing results of Collins (1999)

- Showed that implementing Collins' model using only the published details caused an 11% increase in relative error over Collins' own published results

Fokkens et al., 2013

Discovered five aspects that cause experimental variation not typically described in publications:

- Preprocessing (e.g. tokenization)
- Experimental setup (e.g. splitting data for cross-validation)
- Versioning (e.g. which version of WordNet)
- System output (e.g. the exact features used for individual tokens in NER)
- System variation (e.g. treatment of ties)

Gomes et al., 2018

- *Frustratingly easy domain adaptation* (EasyAdapt) (Daumé III, 2007)
  - Technique that enables developing learning algorithms that perform well across multiple domains
  - **Mixed success** with replication of EasyAdapt in the context of machine translation

Wieling, et al. 2018

- Five studies from each of 2011 and 2016
  - At most 60% reproducible when not enforcing an exact reproduction
  - Exact reproduction only for a single study (from 2011)

Solutions?

GitHub and similar distribution mechanisms

- Easier to distribute versioned code
- **But** many people still report not being able to find code, not being able to remember how to build it, etc.

Maven

- Helps ensure that build processes are repeatable
- **But** most projects in NLP are not distributed as Maven projects
- Maven is not appropriate for every language and architecture used in NLP research

But even given a built program

- May not run due to undocumented platform dependencies, configuration files, input requirements, memory requirements, processor requirements, graphics card requirements, etc.

Solutions

Wieling, et al. 2018

- Provide a virtual (e.g., Docker) image with all software, source code, and data
- Use CodaLab worksheets, Jupyter Notebooks, etc.

Gomes et al., 2018

- Results suggest the importance of **replicating techniques in different contexts** to assess reproducibility of results

Metamorphic testing

- Natural language processing applications are obvious candidates for metamorphic testing
- Applied in situations where we have no “oracle”—situations where we cannot know in advance what the exact output of a function or of a program should be
- General approach
  - Change some aspect of the input for which we can predict in a general way whether or not there should be a change in the output, and what the overall trend in the change should be.
    - E.g., calculate the mean and the standard deviation for some data set, and then add 100 to every value in the data set: mean should increase, standard deviation should not change
    - Examine the stability of performance by running 10 iterations of the largest document set

Activities in the CL Field

- IJCAI workshop on replicability and reproduciblity in NLP in 2015
- Dedicated LREC workshop series “4Real” (workshops in 2016, 2018, and 2020)
- Introduction of a special section of *Language Resources and Evaluation* (Branco et al. 2017) on Reproducibility and Replicability

Summary

- There remains inconsistent use of terminology despite efforts to remedy the situation
  - But there is typically consistency within discipline or area
    - Computational linguistics is an exception!

- The **role that various materials and methods play in different disciplines is key** to providing a truly universal perspective
  - More needs to be done to isolate the relevant variables and apply them appropriately

A Final Word…

- Has all the arguing about which term means what distracted from the main goals?

- Need to focus on:
  - **What is the benefit of the different strategies (exact duplication, duplication minus data, etc.) regardless of terminology?**
  - How will these benefits advance progress?

Thank you!

Bibliography

Branco, A., Bretonnel Cohen, K., Vossen, P., Ide, N., Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. Language Resources and Evaluation, 51(1):1–5.

Broman, Karl, Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D., Wickham, H. (2017).Recommendations to Funding Agencies for Supporting Reproducible Research. American Statistical Association. http://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf.

Cassidy, S., Estival, D. (2017). Supporting accessibility and reproducibility in language research in the Alveo virtual laboratory. Computer Speech & Language 45: 375-391.

Claerbout, J. F., and Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. SEG Expanded Abstracts 11, 601–604.

Cohen, K. B., Xia, J., Zweigenbaum, P., Callahan, T. J., Hargraves, O., Goss, F., Ide, N., Névéol, A., Grouin, C. , Hunter, L. E. (2018). Three Dimensions of Reproducibility in Natural Language Processing. International Conference on Language Resources and Evaluation, 2018, 156–165.

Collberg, C., Proebsting, T. A. (2016). Repeatability in Computer Systems Research, Communications of the ACM, 59:3, 62-69.

Crook, S., Davison, A. P., and Plesser, H. E. (2013). Learning from the past: approaches for reproducibility in computational neuroscience, in 20 Years in Computational Neuroscience, ed J. M. Bower (New York, NY: Springer Science+Business Media), 73–102.

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., and Stodden, V. (2009). 15 Years of reproducible research in computational harmonic analysis. Comput. Sci. Eng. 11, 8–18.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science, in Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML (Montreal, QC). Available online at: http://www.site.uottawa.ca/~ cdrummon/pubs/ICMLws09.pdf

FASEB (2016). Enhancing research reproducibility. Federation of American Societies for Experimental Biology, https://www.faseb.org/Portals/2/PDFs/opa/2016/FASEB_Enhancing%20Research%20Reproducibility.pdf.

Fokkens, A., Erp, M., Postma, M., Pedersen, T., Vossen, P., Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. Proceedings of the 51st Annual Meeting of the ACL, 1691–1701, Sofia.

Gentleman, R., Lang, D. T. (2007). Statistical Analyses and Reproducible Research}, Journal of Computational and Graphical Statistics, 16: 1, 1-23.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? Sci. Transl. Med. 8:341ps12.

Hamermesh Daniel S. (2007). Viewpoint: Replication in economics, Canadian Journal of Economics, 40:3, 715-733.

LeVeque, R.J. (2009). Python tools for reproducible research on hyperbolic problems. Computing in Science & Engineering, 19-27.

Liang, P., Jordan, M. I., Klein, D. (2011). Learning dependency-based compositional semantics. In Proceedings of the 49th Annual Meeting of the ACL, 590–599, Portland, OR.

Barba, L. (2018) Terminologies for Reproducible Research. CoRR abs/1802.03311.

Mesirov, J. P. (2010). Accessible Reproducible Research Science: 327

Mieskes, M.. (2017). A quantitative study of data in the NLP community. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 23–29, Valencia.

Miller, J. N., and Miller, J. C. (2000). Statistics and Chemometrics for Analytical Chemistry. 4th Edn. Harlow: Pearson.

Neveol, A., Cohen, K., Grouin, C., Aude, R. (2016). Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task, Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016, Auxtin, TX, 7884.

Pedersen, T. (2008). Empiricism is not a matter of faith. Computational Linguistics, 34(3):465–470.

Peng, R.D. (2011). Reproducible research in computational science, Science 334 , 1226–1227

Plesser , H. E. (2017). Reproducibility vs. Replicability: A Brief History of a Confused Terminology, Frontiers in Neuroinformatics 11.

Rougier, N. P., Hinsen, K., . Alexandre, F., Arildsen, T. , Barba, L. A. , Benureau, F. C. , Brown, C. T. , De Buyl, P. , Caglayan, O. , Davison, A. P. , et al., (2017☺. Sustainable computational science: the rescience initiative. PeerJ Computer Science, 3, e142,

Schrödinger, E. (1915). Zur Theorie der Fall- und Steigversuche an Teilchen mit Brownscher Bewegung. Physik. Z. 16, 289–295.

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. IMS Bulletin, http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research

Stodden, V., Leisch, F., Peng, R. D. (2014) Implementing Reproducible Research, CRC Press.

Taylor, B.N. and Kuyatt, C.E. (1994) NIST Technical Note 1297, Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results.

Vandewalle, P., Vetterli, M., Kovacevic, J. (2009). Reproducible Research in Signal Processing, IEEE Signal Processing Magazine 26(3):37 – 47.

Wieling, M., Rawee, J., van Noord, G. (2018) . Reproducibility in Computational Linguistics: Are We Willing to Share? Computational Linguistics , 44:4, 642-49.