Literatur lesen und verstehen

HS Experimentelles Arbeiten in der Sprachverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

20. Oktober 2022

Pro/Contra

Wissenschaftliche Beiträge sollten so geschrieben werden, dass eine breite Öffentlichkeit sie versteht.

Pro/Contra

Wissenschaftliche Beiträge sollten so geschrieben werden, dass eine breite Öffentlichkeit sie versteht.

Ablauf

1. Je zwei Gruppen sammeln Argumente für Pro/Contra
2. Argumente werden anschließend im Plenum vorgestellt

Pro/Contra

Wissenschaftliche Beiträge sollten so geschrieben werden, dass eine breite Öffentlichkeit sie versteht.

Ablauf

1. Je zwei Gruppen sammeln Argumente für Pro/Contra
2. Argumente werden anschließend im Plenum vorgestellt

Today

- What is scientific literature and how is it created?
- How to read scientific (NLP) literature

Section 1

Scientific Literature

Computational Linguistics Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Strongly in the CS field

Computational Linguistics Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Strongly in the CS field

Core Requirements for Scientific Literature

- Quality assurance

Computational Linguistics Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Strongly in the CS field

Core Requirements for Scientific Literature

- Quality assurance
  - Reviewing

Computational Linguistics Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Strongly in the CS field

Core Requirements for Scientific Literature

- Quality assurance
  - Reviewing
- Sustainability and (in principle) accessibility
  - It should be possible to access a work in the distant future

Computational Linguistics Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Strongly in the CS field

Core Requirements for Scientific Literature

- Quality assurance
  - Reviewing
- Sustainability and (in principle) accessibility
  - It should be possible to access a work in the distant future
- Publishing houses ensure both (in theory)

Peer Review

- Scientific articles are reviewed by other researchers/scientists
- Blindness
  - Double blind: Reviewer and authors are anonymous
  - Single blind: Only reviewers are anonymous
  - Zero blind / “Open Review”: No one is anonymous
- Different fields have different preferences
  - and different people have different preferences
  - CL: Double-blind (recently reaffirmed)
    - But: Preprint servers are an important venue in machine learning!

Publication Venues

- Monographs (books): Except for theses, typically not reviewed
- Journal articles: Peer reviewed (details are journal-dependent)
- Conference articles: Peer reviewed (details are conference-dependent)
  - "Proceedings" = Collection of all conference articles

Publication Venues

- Monographs (books): Except for theses, typically not reviewed
- Journal articles: Peer reviewed (details are journal-dependent)
- Conference articles: Peer reviewed (details are conference-dependent)
  - "Proceedings" = Collection of all conference articles

Lengths and "Abstracts"

- Length varies
  - Conference articles < 10 pages
  - Journal articles ca. $10 - 50$ pages
- "Abstract"
  - Literal meaning: A summary of an article
  - Conference abstracts (DHd/DH) $\simeq$ short articles

Relevant Publication Venues for CL

Conferences

- ACL / NAACL / EACL / EMNLP: Conferences (double-blind)
  - Association for Computational Linguistics
  - ⚠ ACL 2022: 604 long papers – ACL 2002: 65 papers

aclanthology.org

Relevant Publication Venues for CL

Conferences

- ACL / NAACL / EACL / EMNLP: Conferences (double-blind) aclanthology.org
  - Association for Computational Linguistics
  - ⚠ ACL 2022: 604 long papers – ACL 2002: 65 papers
  - Co-located workshops with more specific focus
    - "Workshop" in CL: Mini conference
  - Workshops associated with *CL conferences also in anthology
- COLING, KONVENS: Smaller conferences

Relevant Publication Venues for CL

Journals

- CL: Uncommon
- Computational Linguistics direct.mit.edu/coli
  - Also in anthology: `https://aclanthology.org/venues/cl/`
  - Fully open access
- Digital Scholarship in the Humanities (Literary and Linguistic Computing) academic.oup.com/dsh
  - Partially open access via UB
- Journal of Computational Literary Studies jcls.io

Relevant Publication Venues for CL

Preprint-Servers

- Origin: Share preprints freely
- No review: Everyone can upload anything
- Popular for machine learning advances
- Many papers are later/also submitted to a conference

arxiv.org

Relevant Publication Venues for CL

Others

- DFG (funding agency): No reviewing → no worth
- Blogs – it depends on their authors
- Sammelbände / collections

Section 2

Reading CL Literature

How to Read?

- Reading scientific literature is work
- A work environment is important
- Reading multiple times is often necessary

How to Read?

- Reading scientific literature is work
- A work environment is important
- Reading multiple times is often necessary

References

- Scientific references consist in:
  - Markers in the text (e. g., “Doe (2015)” oder “[3]”)
  - Bibliographic details at the end
- Different styles
  - CL: author-year
- URLs or DOIs
  - https://www.example.com
  - 10.1515/9783110693973 ⇒ `https://doi.org/10.1515/9783110693973`

Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Axel Pichler/Nils Reiter (2020). "Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003

Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Axel Pichler/Nils Reiter (2020). "Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003

Bei Yu (2014). "Language and gender in Congressional speech". In: *Literary and Linguistic Computing* 29.1, pp. 118–132. DOI: 10.1093/llc/fqs073. eprint: http://llc.oxfordjournals.org/content/29/1/118.full.pdf+html. URL: http://llc.oxfordjournals.org/content/29/1/118.abstract

Hints

- Skim first, read closely later
  - Understand big picture first, details later
- Read selectively (abstract, conclusions, method, ...)
- Take notes
- Read in useful order
- Don't read in one sitting

Hints

- Skim first, read closely later
  - Understand big picture first, details later
- Read selectively (abstract, conclusions, method, ...)
- Take notes
- Read in useful order
- Don't read in one sitting
- Bibliography management tools are convenient to organize notes
  - E.g., zotero, bibtex, citavi

Guiding Questions

You should be able to answer (at least) these questions

- What was the task/the problem to be solved?
- What is the new aspect compared to previous research?
- How well did it work?
  - ⚠ Authors have an interest to highlight success and neglect failure
- Which experiments were made to measure it?
  - Which data and evaluation metrics were used?

Critical Reflection of Literature

- Was there an easier way to achieve similar performance?
- How many assumptions are incorporated (maybe implicit)?
  - What would be needed to redo it from scratch?
  - What would be needed to adapt it to another language/genre/domain?
- Why did the authors did it the way they did?
- Can the experiments actually show what the authors claim they show?
- Are the experiments "correctly" interpreted? Are there alternative interpretations that are just as reasonable?
- Is there evidence to generalize results to "the language", "the text type newspaper", ...?

References I

- Pichler, Axel/Nils Reiter (2020). "Reflektierte Textanalyse". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003.
- Preoţiuc-Pietro, Daniel/Mihaela Gaman/Nikolaos Aletras (2019). "Automatically Identifying Complaints in Social Media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf.
- Yu, Bei (2014). "Language and gender in Congressional speech". In: *Literary and Linguistic Computing* 29.1, pp. 118–132. DOI: 10.1093/llc/fqs073. eprint: http://llc.oxfordjournals.org/content/29/1/118.full.pdf+html. URL: http://llc.oxfordjournals.org/content/29/1/118.abstract.