Lesen und Verstehen computerlinguistischer Literatur

HS Rankingaufgaben in der Computerlinguistik

Nils Reiter
nils.reiter@uni-koeln.de

Department of Digital Humanities

April 13, 2023
(Sommersemester 2023)

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Welche Erfahrungen mit wissenschaftlicher Literatur haben Sie schon gemacht?

Section 1

Overview

Computational Linguistics and its Literature

- Computational Linguistics (CL): A young field
  - Compared to philosophy, physics, ...
- Interdisciplinary between computer science and linguistics
  - Pendular movement
  - Currently: Many influences from computer science

Core Requirements for Scientific Literature

1. Quality assurance

Core Requirements for Scientific Literature

1. Quality assurance: Reviewing

Core Requirements for Scientific Literature

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

Core Requirements for Scientific Literature

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

Core Requirements for Scientific Literature

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

- Publishing houses ensure both (in theory)

"Scientific publishing" $\neq$ making something available to others

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)
  - 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

aclanthology.org

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

Citability

- Some sources are considered “not citable”
  - I.e., they should not be used in scientific references

Citability

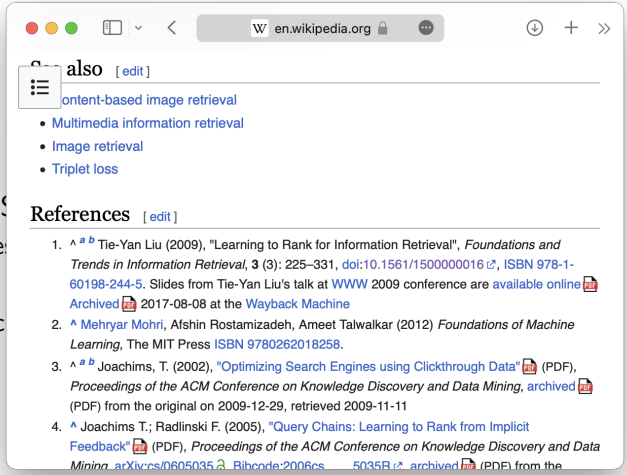
- Some sources are considered “not citable”
  - I.e., they should not be used in scientific references
- YouTube-Videos (no reviewing, no archiving): Not citable
- Company web pages (no reviewing, no archiving): Not citable
  - In most cases written by the PR department anyway
- Blogs (no reviewing, no archiving): Not citable
  - Exceptions apply
- Wikipedia articles
  - Depends on topic and discipline
- Public newspapers / popular science magazines
  - Cite only if absolutely necessary

Finding (Good) CL Literature

- ACL Anthology: Collection of CL publications since 1965 — ACL Anthology
  - Open access
- Google Scholar: Search engine for scientific articles — Google Scholar
  - ⚠ No guarantees on quality
- Wikipedia
  - Articles on scientific topics often have references at the bottom

Finding (Good) CL Literature

- ACL Anthology:
  - Open access
- Google Scholar: S
  - No guarantees
- Wikipedia
  - Articles on sc



ACL Anthology

Google Scholar

Finding (Good) CL Literature

- ACL Anthology: Collection of CL publications since 1965 — ACL Anthology
  - Open access
- Google Scholar: Search engine for scientific articles — Google Scholar
  - ⚠ No guarantees on quality
- Wikipedia
  - Articles on scientific topics often have references at the bottom
- Other articles
  - Related-work-sections show other relevant articles

Structure of a CL Paper

Common structure

- Introduction
- Background
  - Optional. What do we have to know about the phenomenon?
- Related Work
  - Work dealing with same or similar problem
- Approach (the core)
  - Description on conceptual level
  - Good: Point out assumptions the approach makes
- Data set(s) / Corpus
  - Inter-Annotator agreement
- Experiments
  - Baseline(s)
  - Evaluation Metric(s)
- Results
- Error Analysis
  - Types of errors the system makes
- Conclusions
  - Summary
  - Findings about concept(s)
  - Future work

Section 2

Reading (CL) Literature

Was würden Sie Erstsemester:innen zum Umgang mit wissenschaftlicher Literatur raten?

Basics

- Reading scientific literature is work
- A work environment is important
- Reading multiple times is usually necessary
- Scientific literature written to transmit certain ideas within a scientific community

Scientific References

- Important part of scientific writing
- Scientific references consist in:
  - Markers in the text (e. g., “Doe (2015)” or “[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

Bibliographic details

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

Bibliographic details

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

Bibliographic details

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
- Don't read in one sitting
- Read in useful order

Hints

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

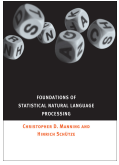
- Was there an easier way to achieve similar performance?
- How many assumptions are incorporated (maybe implicit)?
  - What would be needed to redo it from scratch?
  - What would be needed to adapt it to another language/genre/domain?
- Why did the authors did it the way they did?
- Can the experiments 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 X", ...?

Background Reading

Dan Jurafsky/James H. Martin (2023). *Speech and Language Processing*. 3rd ed. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/

Background Reading

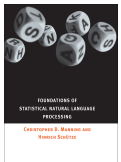
Dan Jurafsky/James H. Martin (2023). *Speech and Language Processing*. 3rd ed. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/



Christopher D. Manning/Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press

Background Reading

Dan Jurafsky/James H. Martin (2023). *Speech and Language Processing*. 3rd ed. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/



Christopher D. Manning/Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press

Hang Li (2014). *Learning to Rank for Information Retrieval and Natural Language Processing*. Ed. by Graeme Hirst. 2nd ed. Synthesis Lectures on Human Language Technologies. Morgan & Claypool

Questions?

References I

- Jurafsky, Dan/James H. Martin (2023). *Speech and Language Processing*. 3rd ed. Draft of Janaury 7, 2023. Prentice Hall. URL: https://web.stanford.edu/~jurafsky/slp3/.
- Li, Hang (2014). *Learning to Rank for Information Retrieval and Natural Language Processing*. Ed. by Graeme Hirst. 2nd ed. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.
- Manning, Christopher D./Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.
- 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.

References II

- 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.