Responsible NLP in the making: contributions from ethics and reproducibility

Aurélie Névéol

LISN (formerly, LIMSI) CNRS

HealTAC, June 18 2021
Acknowledgements

• In 2020, ACL adopted the ACM code of ethics
  – I was a member of the EMNLP 2020, NAACL 2021 and ACL 2021 ethics review committees

• Long standing interest in reproducibility in clinical NLP
  – Organisation of shared tasks
  – Survey and literature studies for a better understanding

• (lack of) accessibility of clinical corpus in French

Bender EM. Academic freedom, academic integrity, and ethical review in NLP. Medium blog post 2021
Ethical considerations
Responsible science:

gatekeeping vs. benevolent guidance
BioNLP research process

- Application
- Community feedback

- Task definition
- Modeling in corpus: Train/test data

- Design, apply NLP methods
- Evaluate using appropriate metrics

- Consider Ethics
- Ensure reproducibility
- Evaluate impact
Ethics is a nuanced discussion of (at least) three aspects of research problems intended to help us strive for better research
Is the problem meaningful and well designed?

Experiment:
Use a cloud hosted language model to test GPT3’s ability to provide:
- Admin chat (e.g. appointment taking) with patient
- Mental health support
- Medical diagnosis
- ...

Human subjects

• When and how to interact with ethics committees, IRB?
• Language involves humans
  • Corpus sources → privacy, potential harm, including essentializing identity characteristics
  • Corpus processing → fair treatment of research participants
  • Corpus downstream users → direct and indirect


How are data and methods designed?

Bias: imbalanced data, diverging data processing, bias amplification


Datasets and corpus development

• Provenance and availability
• Terms of use, including confidentiality, copyrights
  • Some information is always sensitive (e.g. health, religion)
• Detailed description
  • Language, volume
  • Selection and collection method
  • Quality assessment, including biases

Impact of data on evaluation?

• Similarity between training and test corpus
  • 4 biomedical English benchmark datasets
  • Compare performance in redundant vs. non redundant scenarios

• Characterization of memorization vs. generalization
  • What is realistic in a real-life setting?

What is the impact of deployment?

Dr Murphy (aka David Watkins)
@DrMurphy11

Is this another negligent #Triage from the @babylonhealth #G Patt Hand #AI #Chatbot App?

48yr old obese 30/day male smoker develops sudden onset central chest pain & sweating....

I say call 999, the Babylon App says see your GP...
Environmental impact

<table>
<thead>
<tr>
<th>Consumption</th>
<th>CO₂e (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air travel, 1 person, NY↔SF</td>
<td>1984</td>
</tr>
<tr>
<td>Human life, avg, 1 year</td>
<td>11,023</td>
</tr>
<tr>
<td>American life, avg, 1 year</td>
<td>36,156</td>
</tr>
<tr>
<td>Car, avg incl. fuel, 1 lifetime</td>
<td>126,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training one model (GPU)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP pipeline (parsing, SRL)</td>
<td>39</td>
</tr>
<tr>
<td>w/ tuning &amp; experiments</td>
<td>78,468</td>
</tr>
<tr>
<td>Transformer (big)</td>
<td>192</td>
</tr>
<tr>
<td>w/ neural arch. search</td>
<td>626,155</td>
</tr>
</tbody>
</table>

Table 1: Estimated CO₂ emissions from training common NLP models, compared to familiar consumption.¹

Underestimation not accounting for life cycle of computer equipment

Environmental impact

- Reporting of computational resource use
  - Carbon tracker [https://github.com/lfwa/carbontracker](https://github.com/lfwa/carbontracker)

- Benefit/risk analysis beyond leaderboard performance

Reproducibility
Challenges in Reproducibility

• Reports of a reproducibility crisis in many disciplines
  – Poll of 1,500 scientists (2016) 225 NLP researchers (2019)

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Failed to reproduce others’ experiment</th>
<th>Failed to reproduce own experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemistry</td>
<td>90%</td>
<td>60%</td>
</tr>
<tr>
<td>Biology</td>
<td>80%</td>
<td>60%</td>
</tr>
<tr>
<td>Physics and engineering</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>Medicine</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>Earth and environment science</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>Other</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>60%</td>
<td>30%</td>
</tr>
</tbody>
</table>


## Types of reproducibility and gains

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Software/parameters</th>
<th>Data/method</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeat</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Replicate</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Reproduce</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Reuse</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Shared tasks foster reproducibility

- Primary goal is to provide a forum for direct comparison of approaches

- Availability of shared material
  - Specific definition of a “task”
  - Corpora and annotations, split into training, development and test sets
  - Evaluation metrics and scripts
A replicability study at CLEF eHealth 2016 set-up

- As part of the ICD10 coding task, participant could submit their system for replication
  - 4 analysts committed to replicate results in their usual work environment
  - 3 teams submitted systems

- Replication assessment
  - Scoring sheet documenting install/run/results
  - Timing

A replicability study at CLEF eHealth 2016 results

• Results were replicated…
  • No single analyst was able to replicate all
  • Time to replication varied greatly

• But replication is not trivial!

• Replication requires resources
  • For authors to produce quality systems, documentation
  • For others to understand and conduct

A broader look on NLP reproducibility

- Data and code are still elusive


Mieskes M. A quantitative study of data in the NLP community. Proc ACL Workshop on Ethics in NLP. 2017
Towards actionable reproducibility

• From research to hospital operations
  • Need for standardization, traceability, automation

• Leveraging expertise and experience across disciplines
  • Reproducibility criteria/desiderata expressed by the bioinformatics, medical informatics, NLP communities

• Characterize clinical NLP w. r. t. reproducibility
  • Analysis of 7 clinical NLP systems (for English)

40 reproducibility criteria

- R39 Input data publicly available
- R40 Resources publicly available
- R28 Absence of manual steps
- R30 Ability to resume workflow
- R01 Provenance Metadata
- R03 System Metadata
- R04 Record Parameters
- R06 Pipeline versioning
- R07 Tool versioning
- R08 Resource versioning

Evaluation of clinical NLP systems

NLP systems
Relies on UIMA or Gate
- cTakes\(^1\) 18/40
- CLAMP\(^2\) 17/40
- GATE\(^3\) 17/40

WMS Systems
Relies on Galaxy
- LAPPGrid\(^4\) 26/40
- OpenMinTed\(^5\) 22/40
- Textflows\(^6\) 17/40

NLP toolbox python library
- ScispaCy\(^7\) 17/40

• Reproducibility can be improved
  • Especially versioning, standardization and shareability
  • Experience from bioinformatics suggests modularity and workflows can help

Workflow management for French Clinical NLP?

Allows the implementation of a simple workflow integrating “annotators” and text format conversion

https://github.com/equipe22/pymedext_core
Shared clinical corpus in languages other than English
Few corpora are available

- **Death certificates**
  - French, Hungarian, Italian [Névéol et al. 2018]

- **Patient referrals**
  - Spanish [Báez et al. 2020]

- **Creative solutions**
  - Synthetic clinical narratives (Japanese [Aramaki et al. 2014], Norwegian [Rama et al. 2018])
  - Clinical case reports in French [Cardon et al. 2018] and Spanish [Miranda Escalada et al. 2020]
De-identification: a solved problem?

- Research stimulated by shared tasks

- Adressed in several languages
  - English, French, Swedish…

- Still different from anonymization
Synthetic corpus

- Successful attempts for English using neural models
  [Melamud and Shivade 2019, Ive et al. 2020]
  - Trained on MIMIC III
  - Keyphrases and ICD10 codes used as prompts

- Synthetic text visibly different from real corpus but beneficial as data augmentation for processing real data

- Is synthetic data anonymous?
  - Generative models can be tuned for privacy
  - Few rare n-grams retained from original corpus
Summary

• Ethics offers guidelines to question research
  – motives
  – methods and data (including evaluation)
  – impact

• Reproducibility is complex and hard to achieve
  – But we have to keep trying!
  – And document…

• We need to broaden the scope of shareable clinical corpus

Thank you!