Two existential threats to biomedical natural language processing…and how to fight them with biomedical natural language processing

Kevin Bretonnel Cohen
Director, Biomedical Text Mining Group,
University of Colorado School of Medicine;
Emeritus D’Alembert Chair in Natural Language Processing for the Biomedical Domain, Université Paris-Saclay

kevin.cohen@gmail.com
http://compbio.ucdenver.edu/Hunter_lab/Cohen
Objectivity in science is not about not having biases—it’s about being clear about them

- Director, **Biomedical Text Mining** Group, University of Colorado School of Medicine
- Emeritus D’Alembert Chair in **Natural Language Processing for the Biomedical Domain**, Université Paris-Saclay
- Chair, Association for Computational Linguistics Special Interest Group on **Biomedical Natural Language Processing**
  - *Dina does all of the important stuff*
- 150+ publications, mostly in **biomedical natural language processing**
- Cohen and Demner-Fushman 2014, **Biomedical natural language processing**
“Existential threat” defined

**Existential threat**: Something that endangers the continued existence of something else.

*If ambient oxygen less than 21% then living thing might die*

*Ambient oxygen less than 21%*

*…therefore living thing might die*
Existential threat #1: A false syllogism

Syllogism:

- If stuff is good then we should mine it
- Scientific literature is full of good stuff

...therefore we should mine the scientific literature.
...but, the scientific literature might not be full of good stuff

**RIGOR MORTIS**

How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions

RICHARD HARRIS

IS THERE A REPRODUCIBILITY CRISIS?

1,576 researchers surveyed

- 52% Yes, a significant crisis
- 7% Don’t know
- 3% Yes, a slight crisis
- 38% No, there is no crisis

Only **36%** of studies replicated!!

The reproducibility crisis
Erosion of public trust in science puts my grandchildren in danger

Healthy young child goes to doctor, gets pumped with massive shot of many vaccines, doesn't feel good and changes - AUTISM. Many such cases!

5:35 AM - 28 Mar 2014
Erosion of public trust in science puts my grandchildren in danger

Donald J. Trump  
@realDonaldTrump

Give me clean, beautiful and healthy air - not the same old climate change (global warming) bullshit! I am tired of hearing this nonsense.

1:44 AM - 29 Jan 2014
Solution to Threat #1: Use natural language processing to assess rigor

**Science**

Only 36% of studies replicated!!

**Biomedical text mining for research rigor and integrity: tasks, challenges, directions**

Halil Kliceroglu

Corresponding author: Halil Klicative, U.S. National Library of Medicine, Bethesda, MD 20894, USA. Tel.: +1(301)402-7014, Fax.: +1(301)496-2072, E-mail: kklicoglu@narrl.mlb.gov

The reproducibility crisis

*The Economist*

Problems with scientific research

How science goes wrong
"Bag of words" is not enough for strength of evidence classification.

Lin J\textsuperscript{1}, Demner-Fushman D.


Towards automatic recognition of scientifically rigorous clinical research evidence.

This is not a crazy idea.

Toward automatic recognition of high quality clinical evidence.

Kilicoglu H\textsuperscript{1}, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB.


Automated classification of eligibility criteria to facilitate patient-trial matching for specific patients.

Zhang K\textsuperscript{1}, Demner-Fushman D\textsuperscript{2}.
"Bag of words" is not enough for strength of evidence classification.

Lin J\textsuperscript{1}, Demner-Fushman D.


Towards automatic recognition of scientifically rigorous clinical research evidence.

Kilicoglu H\textsuperscript{1}, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB.


Toward automatic recognition of high quality clinical evidence.

Kilicoglu H\textsuperscript{1}, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB.


Automated classification of eligibility criteria to facilitate patient-trial matching for specific patient populations.

Zhang K\textsuperscript{1}, Demner-Fushman D\textsuperscript{2}. 
This is not a crazy idea II: Some preliminary results on NLP for reproducibility

• Cases and controls: kappa = 0.86
• NER of p-values: out-recalls statcheck
• “General science” templates: NIH
• Field-specific templates: autophagy, spinal cord injury

<table>
<thead>
<tr>
<th></th>
<th>Retracted papers versus controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-values</td>
<td>&gt;</td>
</tr>
<tr>
<td>Hedging</td>
<td>&lt; *</td>
</tr>
<tr>
<td>Observer blinding</td>
<td>=</td>
</tr>
<tr>
<td>Multiple testing</td>
<td>=</td>
</tr>
<tr>
<td>corrections</td>
<td></td>
</tr>
</tbody>
</table>

*Hedging finding contradicts previous work—clearly more to be done here*

Thanks to Prabha Yadav and Bill Baumgartner
This is not a crazy idea II: Some preliminary results on NLP for reproducibility

- **Task**: Classify retracted papers
- **Data**: 100 retracted papers, 100 controls
- **Method**: Bag of words, three classifiers
- **Result**: Best F-measure = 0.85
This is not a crazy idea II: Some preliminary results on NLP for reproducibility

- **Task**: Classify retracted papers
- **Data**: 100 retracted papers, 100 controls
- **Method**: Bag of words, three classifiers
- **Result**: Best F-measure = 0.85
This is not a crazy idea II: Some preliminary results on NLP for reproducibility

- **Task:** Classify retracted papers
- **Data:** 100 retracted papers, 100 controls
- **Method:** Bag of words, three classifiers
- **Result:** Best F-measure \(= 0.85\)

- 100 retracted papers, extracted from PDFs
- 100 controls, text extracted from PubMedCentral XML
EXISTENTIAL THREAT #2: WE ARE PROBABLY IN AS MUCH TROUBLE AS EVERYONE ELSE
Is our work generally replicable? No.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data</th>
<th>Sampling</th>
<th>Split</th>
<th>Algorithm</th>
<th>Package</th>
<th>Tokenization selection</th>
<th>System</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen et al. (2016)</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Is our work generally replicable? No.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data</th>
<th>Sampling</th>
<th>Split</th>
<th>Algorithm</th>
<th>Package</th>
<th>Feature selection</th>
<th>System</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen et al. (2016)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Things that **surprised** me: system not available for any of my **machine learning** papers

Totals: 2 6 1 6 3 2 3 0

*Similar results from Aurélie Névéol*
It’s not just me

<table>
<thead>
<tr>
<th>Field</th>
<th>Code</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer science</td>
<td>Kovacevic (2007)</td>
<td>0%</td>
</tr>
<tr>
<td>Computer science</td>
<td>Vandewalle (2009)</td>
<td>9%</td>
</tr>
<tr>
<td>Computer science</td>
<td>Stodden (2010)</td>
<td>30%</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>Cohen et al. (rejected x 2)</td>
<td>7%</td>
</tr>
<tr>
<td>Biomedical natural language processing</td>
<td>Cohen et al. (rejected x 2)</td>
<td>43%</td>
</tr>
<tr>
<td>Me</td>
<td>Cohen et al. (rejected x 2)</td>
<td>0%</td>
</tr>
</tbody>
</table>

*Please ask me about the reviews of Cohen et al. (rejected x 2)*
Are our results typically reproducible? No.

- ludicrous: stupid or unreasonable and deserving to be laughed at (https://dictionary.cambridge.org/dictionary/english/ludicrous)

- ludicrous: amusing or laughable through obvious absurdity, incongruity, exaggeration, or eccentricity (https://www.merriam-webster.com/dictionary/ludicrous)
Why? Natural language processing research is rife with confounds

The results show that a small change in tokenization strategy can improve a mediocre 2006 TREC genomics submission (MAP average: 29%) to the top quarter of the submissions (36%-54%). Normalization and splitting compounds to multiple terms shows to be very beneficial for the tested IR models which assume term independence in both queries and documents. We expect that incorporation of proximities of related terms in the retrieval model will even further improve retrieval performance.

<table>
<thead>
<tr>
<th>POSITIVE INSTANCES</th>
<th>NEGATIVE INSTANCES</th>
<th>F-MEASURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>0.82 ± 0.03</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>0.80 ± 0.03</td>
</tr>
<tr>
<td>100</td>
<td>300</td>
<td>0.74 ± 0.04</td>
</tr>
<tr>
<td>100</td>
<td>400</td>
<td>0.70 ± 0.04</td>
</tr>
</tbody>
</table>

Figure 1. Learning Curves for Confusion Set Disambiguation
HOW CAN NATURAL LANGUAGE PROCESSING FIX THE REPRODUCIBILITY CRISIS WHILE IT’S HAVING ITS OWN REPRODUCIBILITY CRISIS?
The story so far…

• The Reproducibility Crisis
• Text mining is a possible solution
• Computational sciences are probably looking at their own crises…
• …and so is natural language processing
• But: there are actions that we can take at the institutional and individual levels.
LET’S USE BIOMEDICAL NATURAL LANGUAGE PROCESSING TO FIX BIOMEDICAL NATURAL LANGUAGE PROCESSING

(...and everything else, too)
The closer we look at natural language processing, the more problems we see... So: lots of research opportunities

- What we have been thinking about: data (Mieskes 2017), code (Pedersen 2008)
- What we have not been thinking about (much): Experimentation
  - “NLP-hacking” (Cohen et al. 2019)
  - Small differences that could easily be noise
  - Bugs (Cohen et al. 2007-2018)
  - **Confounds**... But how do we know what they are?
Plausible things to do at the institutional level

**Funding agencies**: “key computational variables”

**Conferences/journals**: review criteria
- Replicability criterion *Ask me about this later if you would like details*
- Reconsider “innovation” *This, too*
- **Value experimental validity over (putative) performance**
**Funding agency level: develop the “key computational variable”**

| Biological Variables | Biological variables, such as sex, age, weight, and underlying health conditions, are often critical factors affecting health or disease. In particular, sex is a biological variable that is frequently ignored in animal study designs and analyses, leading to an incomplete understanding of potential sex-based differences in basic biological function, disease processes and treatment response. Explain how relevant biological variables, such as the ones noted above, are factored into research designs, analyses, and reporting in vertebrate animal and human studies. Strong justification from the scientific literature, preliminary data or other relevant considerations must be provided for applications proposing to study only one sex. |

How do you know what is a “key computational variable”?

Model on definition of key biological variable?

- NIH FAQ on Rigor and Transparency
- NIH Center for Scientific Review guidelines on Premise, Rigor, and Sex as a Biological Variable
- NIH Reviewer Guidance to Evaluate Sex as a Biological Variable (SABV)
How do you know what is a “key computational variable”?

Model on definition of key biological variable?

- NIH FAQ on Rigor and Transparency
- NIH Center for Scientific Review guidelines on Premise, Rigor, and Sex as a Biological Variable
- NIH Reviewer Guidance to Evaluate Sex as a Biological Variable (SABV)
Clayton (2018): Literature search is the first step

5. How can SABV be implemented in research and in clinical practice?

5.1. Research

Considering SABV is not the same as looking for sex differences, but it is about exploring the influences of sex as a biological variable and revealing the “data hiding in plain sight.” Here, we describe some ways to help researchers “see” potential influences of sex. First, before conducting research, find out whether there are known sex differences in the area of study by adding the terms “sex,” “gender,” “male,” and “female” to your literature search. In addition to PubMed, use the GenderMed database. Second, randomize and balance the sexes in the
This is not crazy III

- Gricean maxim of relevance: assume that if someone is talking about something, it “matters”

If it is in the title of a paper, it matters.

X is in the title of a paper.

X matters.
5. How can SABV be implemented in research and in clinical practice?

5.1. Research

Considering SABV is not the same as looking for sex differences, but it is about exploring the influences of sex as a biological variable and revealing the “data hiding in plain sight.” Here, we describe some ways to help researchers “see” potential influences of sex. First, before conducting research, find out whether there are known sex differences in the area of study by adding the terms “sex,” “gender,” “male,” and “female” to your literature search. In addition to PubMed, use the GenderMed database. Second, randomize and balance the sexes in the
2 repositories, several “levels” of task: complicated, but works

<table>
<thead>
<tr>
<th>Preprocessing</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization</td>
<td>3</td>
<td>(“natural language processing&quot; OR &quot;text mining&quot;) (tokenization[ti] OR tokenizer[ti])</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>alltitle: tokenizer OR tokenization OR tokenizing source:computational source:linguistics</td>
</tr>
<tr>
<td>Stemming</td>
<td>2</td>
<td>(“natural language processing&quot; OR &quot;text mining&quot;) (stemming[ti] OR stemmer[ti])</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>alltitle: stemming OR stemmer source:computational source:linguistics</td>
</tr>
<tr>
<td>POS tagging</td>
<td>15</td>
<td>(“natural language processing&quot; OR &quot;text mining&quot;) (POS[ti] OR &quot;part of speech&quot;[ti])</td>
</tr>
<tr>
<td></td>
<td>195</td>
<td>alltitle: &quot;part of speech&quot; source:computational source:linguistics</td>
</tr>
<tr>
<td>NER</td>
<td>9</td>
<td>(“natural language processing&quot; OR &quot;text mining&quot;) (&quot;information extraction&quot;[ti]) &quot;named entity recognition&quot;[tiab]</td>
</tr>
</tbody>
</table>

| Parsing: structural | 44| (“natural language processing" OR "text mining") (parser[ti] OR parsing[ti]) NB: could include some on dependency parsing |
|                     | 1780| alltitle: parser OR parsing source:computational source:linguistics |
|                     | 1370| alltitle: parser OR parsing -dependency source:computational source:linguistics |
| Parsing: dependency | 20| (“natural language processing" OR "text mining") (dependencies[ti] OR dependency[ti]) |
|                     | 854| alltitle: dependency OR dependencies source:computational source:linguistics |
| Feature selection   | 14| (“natural language processing" OR "text mining") "feature selection"[ti] |
|                     | 38| alltitle: "feature selection" source:computational source:linguistics |
| Smoothing           | 3 | (“natural language processing" OR "text mining") smoothing[ti] |
|                     | 21| alltitle: smoothing source:computational source:linguistics |
Searching in titles for things that we usually do not document: 4 examples

**Tokenization**


**Feature selection**


**Smoothing**


**Back-off mechanism**

Yes, we can automate this

```r
#write to file
write.csv(topic1ToTopic2, file=paste("LDAGibbs", k,"Topic1ToTopic2.csv"))
write.csv(topic2ToTopic3, file=paste("LDAGibbs", k,"Topic2ToTopic3.csv"))

# Validation:
print(paste("Files processed:", counter.files.processed))

## [1] "Files processed: 2393"

print(paste("Files skipped:", counter.files.skipped))

## [1] "Files skipped: 607"

Cute fact: 25% of XML files were invalid

# for replicability:
sessionInfo()

## R version 3.4.1 (2017-06-30)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS 10.14.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
```
What do we find? There are many computational equivalents of “sex”

Strong justification from the scientific literature, preliminary data or other relevant considerations must be provided for applications proposing to study only one sex.

- Computer science
  - Nature of test data (blind, cross-validation, high-frequency only…)
  - Appropriate metric (not accuracy)
  - Appropriate baselines (trivial versus strong)

- Natural language processing
  - Domain
  - Genre
  - Language
  - Evaluation approach/metric
  - Preprocessing approach

There are many natural language processing equivalents of “sex” **in the biomedical domain**

- **Domain**
  - General
  - Scientific
  - Clinical
  - Consumer-generated

- **Genre**
  - Discharge summaries
  - ICU progress notes
  - Death certificates

- **Evaluation approach/metric**
  - Intrinsic, metric, +/- hypothesis testing, …
  - Extrinsic, time, memory, cost, …

- **Language**
  - Analytic (Mandarin)
  - Inflectional (Spanish, French)
  - Synthetic (English)
  - Agglutinative (Finnish, Turkish, Hungarian)
  - Free word order (Russian, Hungarian)
Association/journal level: Prioritize validity over putative performance

**Leaderboard model**

**Deresultified model**


Smarter people than me have given this example—Ken Church, Christopher Manning, Bonnie Webber?
Association/journal level: Prioritize validity over putative performance

**Leaderboard model**

Table 5. Classification using Gene Ontology concepts
Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>GOA curated P/R/F</th>
<th>NLP abstracts P/R/F</th>
<th>NLP full-text P/R/F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unbalanced Training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random OneR</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMO</td>
<td>0.66602</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>LibSVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Balanced Training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random OneR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random Forest</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SMO</td>
<td>0.64644</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>LibSVM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


**Deresultified model**

Smarter people than me have given this example—Ken Church, Christopher Manning, Bonnie Webber?
Back to Threat #1: Most NLP tasks are relevant to commonly cited causes of reproducibility problems

- **Named entity recognition**
  - NER: cases, controls, …
  - Relation extraction: count to class
  - Data set construction: REALLY hard
  - Annotation

- **Relation extraction**
  - NER of p-values
  - NER of multiple testing corrections
  - Relation extraction: test to p-value to result

---

...and plagiarism detection, and hedging detection, and deception detection, and…

- **Spin detection:**


Back to Threat #1: Some preliminary results on NLP for reproducibility

- Cases and controls: kappa = 0.86
- NER of p-values: out-recalls statcheck
- “general science” templates:
- Field-specific templates: autophagy, spinal cord injury

<table>
<thead>
<tr>
<th>Retracted papers versus controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-values</td>
</tr>
<tr>
<td>Hedging</td>
</tr>
<tr>
<td>Observer blinding</td>
</tr>
<tr>
<td>Multiple testing corrections</td>
</tr>
</tbody>
</table>

* Hedging finding contradicts previous work—clearly more to be done here
Why biomedical natural language processing?

- Scientific literature is full of good stuff
- If stuff is good then we should mine it

...therefore we should mine the scientific literature

Gordin, Michael D. *Scientific Babel: How science was done before and after global English*. University of Chicago Press, 2015.
Things that we can do at the individual level

- Open Access code and data (Pedersen 2008)
- Software testing (Cohen et al. 2004-2010, Demner-Fushman 2008-2014)
- Everything that was supposed to be a silver bullet, but wasn’t (but clearly helps)
  - Version control
  - Markdown
- Experimental design
- **Training**
For example…

- An annotation project on negation in translational research (bench-to-bedside)
  - Why this: I really, really tried to make this replicatable, so if this weren’t replicatable, it would mean something serious
  - I had a bunch of students who needed to do an annotation project
Hi, folks,

I just committed the two files that I know were missing. Anything else that you find missing, that is not easy to figure out, etc.--please mail the rest of us about it! I do want to know what stands in the way of you being able to repeat what I did.

Kev
Hi Kevin-

Trying to work on this now. I cannot download the PDF of the original paper stored on GitHub. An error message saying the file is damaged keeps popping up.
3 and 4: data files not in repository or not named as expected

- I noticed that there is a document that you reference in the paper that is not in the repo (*negativesEvery10KWords.pl*).

- In running through the .Rmd file, I had to modify some of the file names it listed in order to run the code. Specifically, the following extensions had to be modified to make what was in the repo:
  - `CRAFT.type.token.with.prefix.KEVIN.csv` —  
    > `CRAFT.type.token.with.prefix.KEVIN.txt`
  - `MIMIC2MD.type.token.with.prefix.KEVIN.annotated.csv` —  
    > `MIMIC2MD.type.token.with.prefix.KEVIN.annotated.xls`
5: file formats don’t match I/O requirements of the script

- If you make these changes then additional changes are needed in order to read the files in correctly without error. For instance, changing comma separated to tab delimited. With these changes the markdown file runs without any problems and I get the same number that you report in the manuscript
  - I made these changes and submitted a pull request. I did not re-generate the PDF form this Markdown file as doing that would require me to alter the working directory, but the .Rmd file has been updated.
6: more missing files (to run a script that I should not have put in the repository!)

- The `createExperimentDataSet.sh` bash script requires a bunch of files that are not in the GitHub directory so I get a bunch of errors when trying to run it:
  - `CRAFT.type.token.txt`
  - `MIMIC2MD.type.token.txt`
  - `distractors.nonbiomedical.multimorphemic.csv`
  - `distractors.not.ambiguous.biomedical.csv`
  - `distractors.not.ambiguous.nonbiomedical.csv`
  - `negatives.wikipedia.plain.english.csv`
...and then my reward!

Hey Kevin-

I am working on calculating the IRR for my annotations now and noticed something in the code I wanted to ask you about. See snippet below:

```r
agreement.craft <- agreement(mimic.kevin.data, mimic.fiji.data)
print("CRAFT IAA (only including ones marked by someone):")
agreement.craft

agreement.mimic <- agreement(craft.kevin.data, craft.fiji.data)
print("MIMIC IAA (as above):")
agreement.mimic

agreement.craft <- agreement.including.all(mimic.kevin.data, mimic.fiji.data)
print("CRAFT IAA (including all):")
agreement.craft

agreement.mimic <- agreement.including.all(craft.kevin.data, craft.fiji.data)
print("MIMIC IAA (as above):")
agreement.mimic
```

It looks like what is labeled as CRAFT was actually run on MIMIC data and not on the CRAFT annotations. Or maybe I am missing something?
Everything I know, I learned from my students

Tiffany Callahan

Taylor Erwin
Conclusion: Our trainees are the key to fixing this

Source: http://chronicle.uchicago.edu/990415/mccawly.shtml

Picture credit: Hans Peters Anefo, Nationaal Archief (CC BY 4.0)
Thank you colleagues in this work: Bill Baumgartner, Tiffany Callahan, Nicoletta Calzolari, Taylor Erwin, Karën Fort, Cyril Grouin, Larry Hunter, Helen Johnson, Margot Mieskes, Aurélie Névéol, Anna Ripple, Jingbo Xia, Pierre Zweigenbaum, ...  

Acknowledgments
- Funding from the Biomedical Informatics Training Program
- Funding from administrative supplement to Computational Bioscience Program NLM T15
- Dina Demner-Fushman, Larry Hunter, Yassine Mrabet, and John Wilbur—for disagreeing
How bad is it in computational sciences? Low incidence of availability, buildability…

- Replication:
  - Data availability
  - Code availability
  - Buildability

- Publication bias hides negative results
We present results on a quantitative analysis of publications in the NLP domain on collecting, publishing and availability of research data. We find that a wide range of publications rely on data crawled from the web, but few give details on how potentially sensitive data was treated. Additionally, we find that while links to repositories of data are given, they often do not work even a short time after publication.


<table>
<thead>
<tr>
<th>Venue</th>
<th># papers</th>
<th># data published</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAACL</td>
<td>182</td>
<td>57</td>
<td>31.3%</td>
</tr>
<tr>
<td>ACL</td>
<td>231</td>
<td>63</td>
<td>27.3%</td>
</tr>
<tr>
<td>EMLNP</td>
<td>264</td>
<td>81</td>
<td>30.7%</td>
</tr>
<tr>
<td>Coling</td>
<td>337</td>
<td>89</td>
<td>26.4%</td>
</tr>
<tr>
<td>LREC</td>
<td>744</td>
<td>414</td>
<td>55.6%</td>
</tr>
<tr>
<td>total</td>
<td>1758</td>
<td>704</td>
<td>40.0%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link available</td>
<td>65.2%</td>
</tr>
<tr>
<td>Link does not work</td>
<td>15.7%</td>
</tr>
<tr>
<td>No Link</td>
<td>31.4%</td>
</tr>
<tr>
<td>On Request</td>
<td>1.8%</td>
</tr>
<tr>
<td>Proprietary data</td>
<td>&lt; 1%</td>
</tr>
</tbody>
</table>

Table 2: Detailed numbers on available and working links.
Association/journal level: Two ways to make replicability a review criterion

Typical drop-down

• 5: I have rerun the entire thing from GitHub and got identical results
• 4: Appendix gives version numbers, parameters, etc.
• 3: …
• 2: …
• 1: I have no idea what was actually done or with what materials

Checklist

- Tool name
- Tool version number
- Tool language model
- Parameter settings
  - gaps
  - wordOrder
  - acronymAbb
Experimental design: Over-optimizing
We can test generalizability of natural language processing results... We just need to control better.
Why biomedical natural language processing?

- The OntoNotes project comprises a number of different annotations of the same text, in different annotation levels. These levels include coreference. The OntoNotes conference annotation differs from most prior projects in that it includes event coreference, which allows verbs to be markables [36]. The OntoNotes guidelines were the primary source of the CRAFT conference annotation guidelines, and OntoNotes will be discussed in more detail below. Version 4.0 of the OntoNotes data was distributed in the context of the CoNLL 2011 shared task on coreference resolution [37].

The significance of the work reported here comes in part from its focus on biomedical literature, as opposed to the large body of previous work on general-domain materials. As discussed elsewhere in this paper, general-domain conference resolution systems have been found to not work well on biomedical scientific publications [22, 23]. This observation holds within a context of widespread differences between biomedical and general-domain text. Biomedical scientific publications have very different properties from newswire text on many linguistic levels, and specifically on many levels with relevance to natural language processing and text mining. Lippincott et al. [18] looked at similarities and differences in a number of linguistic levels of a wide variety of linguistic levels of newswire text and of scientific text in a broad cross-section of biomedical domains, and found that newswire text almost always clustered differently from scientific texts with respect to all linguistic features, including at the morphological level (e.g. distribution of lefics categories and marking of word-internal structural type), at the relational level (e.g. typographic features and lexical category), and at the syntactic level to small differences in tokenization strategies (e.g. the lexical level (e.g. distributional properties of the lexic [18], weaker prediction power of determinists features for named entities, and length distributions of named entities). The syntactic level (e.g. syntactic structures at the outside of the grammar of newswire text, structural benefits in the distribution of syntactic annotations such as transitivity and argument structure) and longer, more complex sentence structures and distributions of determiner-noun phrase as well as longer dependency chains and complex noun phrase length and presumably complexity [11], and the semantic level (e.g. the type and complexity of semantic classes and the relations, and domain-specific patterns of polynomic lower level 91). The power of lexical features in citation encoding (e.g. in the number and gender distribution (and therefore relative usefulness or lack thereof) of number and gender cues in anaphora resolution) [11], the distribution of anaphoric relation types and the presence of named entities versus complex noun phrases as the antecedents of anaphora [59]. Striking differences in the use of cognitively salient forms related to sensory experience and time have been noted between newswire and scientific text, as well as in light of these numerous differences between newswire text and biomedical text at every linguistic level, the differences that have been noted between newswire text and biomedical text are not surprising. They motivate the work described in this paper.

We turn here to the description of number of biomedical conference corpora, about none of which are publicly available, making the significance of the CRAFT conference annotation project clear.

- Enzweiler et al. [6] annotated a set of 100 sentences per year of MedLINE abstracts, finding that the number of anaphora per a text (1.67. In this context, mostly anaphora that refer back to an immediately preceding case category to which they belong e.g. MAPK and MAPK... etc. (the kinases).
- Yang et al. [69] annotated a corpus of 20,000 sentences per year MedLINE abstracts from the MEDIA data (47). They demonstrated that it is possible to annotate all conference in scientific publications. Descriptive statistics of the annotations are given in Table 1 for comparison with the distribution of annotations in the CRAFT conference corpus.
- Kim and Park [60] created a corpus annotated with pronouns, anaphoric noun phrases with determiners, and zero pronouns. The descriptive statistics are given in Table 2.
- Sanchez et al. [64] annotated a corpus consisting of mixed abstracts and full-text journal articles from the MEDLINE text abstracts from the MEDICOREL library and the Journal of Biological Chemistry. A number of interesting findings came from the analysis of this corpus, including that 5% of proteins-protein interaction assertions contain anaphora, with pronominal anaphora outnumbering verbal anaphora by 18 to 1, even though verbal anaphora are more frequent than pronominal anaphora in biomedical texts in general.

Table 1 Descriptive statistics of Yang et al. at a conference corpus [16]

<table>
<thead>
<tr>
<th>Anaphora type</th>
<th>Total number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphora marked</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noun phase</td>
<td>3361</td>
<td>29.1%</td>
</tr>
<tr>
<td>Pronoun</td>
<td>131</td>
<td>1.1%</td>
</tr>
<tr>
<td>Noun-anaphora marked</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noun phase</td>
<td>8272</td>
<td>67.8%</td>
</tr>
<tr>
<td>Pronoun</td>
<td>259</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

A (generalizability) problem exists, and we have plenty of cases.
Why biomedical natural language processing?

- Constrained enough to be tractable
  - Unidirectional
  - Definable test collections

• The OntoNotes project comprises a number of different annotations of the same text, in different annotation levels. These levels include conference. The OntoNotes conference annotation differs from most prior projects in that it includes event conference, which allows verbs to be participial.

The OntoNotes guidelines were the primary source of the CRAFT conference annotation guidelines, and OntoNotes will be discussed in more detail below.

Version 4.0 of the OntoNotes data was distributed in the context of the CoNLL 2011 shared task on conference resolution [37].

The significance of the work reported here comes in part from its focus on biomedical literature, as opposed to the large body of previous work on general-domain material. As discussed elsewhere in this paper, general-domain conference resolution systems have been found to not work well on biomedical scientific publications [22, 23]. This observation holds within a context of widespread differences between biomedical and general-domain text. Biomedical scientific publications have very different properties from newswire text on many linguistic levels, and specifically on many levels with relevance to natural language processing and text mining. Lippincott et al. [16] looked at similarities and differences in a number of linguistic levels of a wide variety of linguistic levels of newswire text and of scientific text in a broad cross-section of biomedical domains, and found that newswire text almost always clustered differently from scientific texts with respect to all linguistic features, including at the morphological level (e.g. distribution of lexical categories and marking of word-internal structure), at the relation between typographic features and lexical category, and at sensitivity to small differences in tokenization strategies. On the syntactic level, robust predictive power of deterministic features for named entities, their length, and length distributions of named entities. The major syntactic level (e.g. syntactic structure) at the outside of the grammar of newswire text. The results in the distribution of sympathetic annotations such as transitivity and transitivity length for longer, more complex sentences and the distribution of semantic and noun phrase length and presumably complexity [17] and the semantic level (e.g. the typicality of complex of semantic classes and the relation of domain-specific patterns of polysemic lower determining power of lexical features in citations encoding prepositional number and gender distribution (and therefore relative usefulness or lack thereof of number and gender cues in anaphora resolution) to the distribution of anaphoric relation types and prevalence of named entities versus complex noun phrases as the antecedents of anaphora [39]. Striking differences in the use of cognitively salient cues related to sensory experience and time have been noted between newswire and scientific text, as well as subtle differences between these categories.

We turn here to the discussion of the number of biomedical conference corpora. Amounts of data are publicly available, making the significance of the CRAFT conference annotation project clear.

• Sato et al. [62] annotated 1993 and 1994 Anaphora in MEDLINE abstracts, finding 2054 total anaphoric noun phrases in 2033 articles (mean: 9.9); 840 instances of pronoun anaphora (mean: 1.9); 976 instances of proper noun anaphora (mean: 4.3); 217 instances of pronominal anaphora (mean: 2.1); and 502 instances of possessive anaphora (mean: 2.5). Sato et al. [62] used the following statistics:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphora</td>
<td>2033</td>
</tr>
<tr>
<td>Pronoun</td>
<td>1993</td>
</tr>
<tr>
<td>Proper Noun</td>
<td>1993</td>
</tr>
<tr>
<td>Possessive Pronoun</td>
<td>1993</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics of anaphora in a conference corpus [20].
Why biomedical natural language processing?

The clinical narrative has unique characteristics that differentiate it from scientific biomedical literature and the general domain, requiring a focused effort around methodologies within the clinical NLP field.

- Scientific literature
- Clinical documents
- Patient-generated text

Reviewer comments (top conference in natural language processing)

- …each annotator didn't annotate all papers provided. Why not? If each annotator annotated all papers, there can be more pairwise agreements.
- …there are some experiment detail errors and written errors. It can be better.
Reviewer comments (top conference in natural language processing)

- text in the caption is already discussed in the main text.
- strange beginning of sentence.
- Lines 45-49 - I would expect to find data in a non-proprietary format.
Reviewer comments (good conference on computational biology)

• ...suffers from a small sample size.
• ... all the results are in tables which aren’t particularly stimulating
• ...the comparison set of papers that win a Best Paper Award may actually be biased in an unknown direction: are they necessarily more reproducible? They may actually be less so since awards are sometimes given to more flashy results
• It would be useful for the authors to add discussion of how their approach might scale.
• mainly qualitative discussion of the comparisons between different domains (e.g. biomedical natural language processing vs. general natural language processing) but little or no formal testing of these claims.
• more as a report than as a scientific article.
• I’m interested in a general analysis of machine learning NLP methods similar to the authors’ own papers in Section 2.4.
• only evaluated using two criteria for reproducibility
• While this is useful for an initial comparison against both CS and NLP fields, this evaluation lacks rigor and requires further evaluation. Reproducibility is not just the availability of data but having all the required data to reproduce the analyses from the original paper. By focusing only on data and code availability, the paper did not fully address the reproducibility problem in the biomedical NLP space.
• comparison between this paper (general and bio NLP) and previous work (CS) assumes there has been no improvement in CS reproducibility. Because of the 7-9 year difference between the current analysis and the original CS papers, there may have been an increase in reproducibility in CS over the last decade but this is unaddressed by the current paper.
• The utility of the qualitative score is not fully explored.