Accelerated Natural Language Processing 2018

Lecture 29: Ethics and Social Responsibility

Henry S. Thompson
20 November 2018

1. Why ethics?

Studying ethics doesn't really seem a necessary part of studying NLP

- So why are we talking about ethics?

Because we suppose you intend to *use* what you're studying here

- And ethics is not just something you can study
- It's something you *do*

So part of knowing *how* to work as an NLP practioner

- Is knowing how to do so ethically
- And acting ethically as a scientist in general
  - and in particular as a scientist engaged with NLP
- involve ethical issues which are arguably different from the 'everyday' ethical issue

We'll look at three levels of ethical issue:

1. Social responsibility as a (computer) scientist
2. Scientific ethics as such
3. Ethical issues more-or-less specific to NLP

2. Social responsibility

Disclaimer: There are many different views about the proper relationship of society, the state and the individual

- in terms among others of the rights and responsibilities of each

What follows express my non-specialist understanding of the Western liberal tradition that I'm a part of

- If you come from a different tradition, some or all of your own views may be quite different

Key starting point: There is no such thing as value-free science

- That is, scientific activity is almost *never* without practical consequences
- So good for society, some not so good

So a scientist has, or at least shares, an ethical *duty* to promote the good consequences and try to forestall the bad ones

3. Public and private action

Groups or individuals may decide the bad uses of their work are bad enough to withdraw altogether

- Edinburgh's Department of Artificial Intelligence (one of the predecessor constitutents of today's School of Informatics) had a policy of not applying for funding from the Ministry of Defense
- Computer Professionals for Social Responsibility (in the USA) and Computing and Social Responsibility (in the UK) promoted and organised widespread opposition among computer scientists opposed to the Strategic Defense Initiative (known as "Star Wars") in the 1980s, *on the grounds that it crucially depended on unachievable contributions from computer science.*
- In the context of the same programme, the British Computer Society supported the right of individual members of the profession to withdraw from projects on ethical grounds
- Edward Snowden resigned from the US National Security Agency (NSA) and subsequently released a large amount of classified information about NSA's surveillience activities

There is a wide range of individual action available

- Withdrawal of participation (**conscientious objection**: *ohne mich*, "count me out")
- Public denunciation (may amount to whistleblowing)
- Organising opposition
- Civil disobedience (ref. Snowden)

4. Scientific ethics as such

Taking responsibility for the consequences of your work is a kind of **extrinsic** matter

- But science has its own **intrinsic** ethics
- Norms that govern how science should be carried out

Parallel to the wider social responsibility level

- Intrinsic *scientific* ethics can be seen as governing the proper relationship between a scientist and the scientific community
- Much of it amounts to different kinds of positive and negative obligations with respect to honest communication:
  - Publish your own work (not other people's)
  - Give credit where credit is due
  - Give full details, so that others can replicate
  - Don't silently edit experimental results
  - Don't selectively report experiments
    - That is, if you run a dozen experiments
      - report on *all* of them
      - not just the ones that 'worked' (remember XKCD's green jellybeans)
    - Particularly relevant for drug trials
    - See Clinical Trial Regulation requirements in the EU
  - Participate in peer review
    - Submit your own work for review
    - Act as a reviewer for the work of others

5. Open Data

The empirical/data-driven methodology of contemporary NLP

- Was fueled by a mind-shift about language data collections (**corpora**)
- Replacing "my scientific identity is defined by my data"
  - so, I keep my data to myself
- with "If we all share our data, we all benefit"
  - because, there's no data like more data
- in a positive-feedback loop (or **virtuous circle**)

This view about data is now much more widespread

- Where it is often referred to as **Open Data**
- See, for example, the FAIR Data principles
- Scientific data should be Findable, Accessible, Interoperable and Reusable

6. Open Science/Open Access

Open Data is about the *input* to scientific work

- What about the *output*?

Historically, good science was understood to mean science that was published in peer-reviewed journals

- The tougher the reviewing process
- The better the science
- And the more expensive subscriptions to the journal

Only big university libraries and big companies could afford to subscribe to a reasonable number of the good journals

The Web has begun to change all that

- The effective monopoly of the major journal publishers is breaking down
- Research funding bodies are starting to require publication in free-to-read-online journals (**Open Access**)

But this has had some negative consequences

- Putting peer review at risk
- The growth of Article Publishing Charges (APCs)

The whole question of dissemination of scientific results has become very complicated

7. What about "404 Not Found"?

Open Access means the publication of record for a rapidly growing number of articles is online

- But exactly *where* online?
- And more to the point, for *how long*?

Even before the Open Access movement **link rot** was becoming a serious problem

- That is, the web links (URIs) in online papers sometimes either
  - Didn't work at all (clicking on them resulted in "404 Not Found") or
  - Worked, but didn't result in the intended content

Publishers reorganise their websites, so articles get new URIs

- Or get taken over, and their archives are moved or (worse) lost

This has led to the rise of third-party vendors of so-called **persistent identifiers** (PIDs)

- Managed spaces of formal identifiers which redirect to wherever an article moves to

The most widely used of these is the Digital Object Identifier (DOI)

- Originally written as e.g. `doi:10.1145/3184558.3191636`
- Now (more usefully) as https://doi.org/10.1145/3184558.3191636

Note that not all PIDs are for articles

- And not all PIDs are **actionable**
- That is, written as `http:` or `https:` URIs

For example, my ORCID is `0000-0001-5490-13`

- That's an Open Researcher Contributor IDentifier
- Which *can* be looked up as https://orcid.org/0000-0001-5490-13

And the database identifier for the human genome is `taxon:9606`

- Which can be looked up as https://identifiers.org/taxon:9606

8. Before NLP-specific ethics, some necessary background

Where does our data come from?

- Historically, collected from non-online sources
  - Distributed on the best bulk-transfer medium of the day

The Brown Corpus was the first attempt (1967) at a machine-readable corpus

- 1 million words of English
  - 500 samples of 2000 words each
- Transcribed onto punch-cards
- Distributed via reels of magnetic tape

The ACL/DCI Wall Street Journal corpus (Association for Computational Linguistics/Data Collection Initiative) (1993) was the first big step upward in scale

- 30 million words
- From the 1987–89 editions of the Wall Street Journal
- Converted from the printers' tapes by Mark Liberman
- Distributed on CD-ROM
- Helped launch the Linguistic Data Consortium (LDC)

The first substantial distributions of non-English data came from Edinburgh

- The first (1994) an *ad-hoc* collection of freely-available data in over 20 different languages (ECI/MC1)
- The second (1997) a more carefully designed collection of text from 6 major European financial newspapers along with parallel material in 9 languages from the Commission of the European Community (MLCC)
- Each containing about 90 million words, both distributed on CD-ROM
- Helped launch the European Language Resources Association (ELRA)

9. Intellectual Property

Your ideas, your writing and your speech are your **intellectual property**

- Does that mean you *own* them?

Yes, in a way

- Language data (text, speech) is not free
- Many (most?) legal jurisdictions recognise two ways in which people *own* language:
  - Trademark
  - Copyright
- Trademarking covers individual words or phrases and is rarely relevant for NLP
- But copyright is a major issue

The details of copyright vary in different legal systems

- But the basic idea is pretty much the same:
  - You control the extent to which your words can be reused
    - Copied
    - Performed
    - Reworked (creating **derivative works**)
  - Reuse without permission may "violate copyright"

In some jurisdictions, notably the United States, some forms of copyright violations are treated as major crimes (felonies)

10. Copyright: some details

Copyright is inherent

- You don't have to register your words
- You don't have to include "Copyright © me 2018"

Copyright expires

- It doesn't last forever (for example 70 years after the author's death, in the EU)

Some kinds of copying are allowed, for example

Fair use/Fair dealing
: The right to extract for purposes of review or scholarship

Comedy
: The right to derive certain types of artistic works: "caricature, parody or pastiche" (EU again)

No harm, no foul

- That is, unless the copyright holder's commercial/reputational/financial interests have been affected, there is no case to answer

Broadcasts

- Copying radio or television broadcasts for private (non-commercial) use is allowed (UK as of 1958)

11. Licensing rights

Copyright holders can **license** their rights

- An author has to license his/her rights to a publisher

Corpus creators have to get licenses

- And grant them in turn to whomever they deliver the corpus to

The vast majority of the work in creating the MLCC was in the licensing negotiations

The reason you can't download the Twitter data for last week's lab

- The terms of the original license from Twitter

The Informatics corpus collection is divided on the basis of license terms

12. Using licensed data for research

It's usually straightforward to follow legal and ethical guidelines when using data under license

- Don't redistribute data without checking license agreements
  - This includes modified or annotated versions of the data
- In most cases, you may store your own copy of data licensed by Edinburgh *to use for University-related work*, but always check the license
- If in doubt, check with your instructor or project supervisor

13. Data from the Web

The Web is itself a corpus

- Not designed or curated
- But *very* big!

By far the largest source of language data now available

- Requires **crawling** the Web to collect pages, and **scraping** them to extract language data

Some notable examples:

Google Ngrams
: "Web 1T 5-gram Version 1, contributed by Google Inc. [to the LDC], contains English word n-grams and their observed frequency counts. The length of the n-grams ranges from unigrams (single words) to five-grams. ... The n-gram counts were generated from approximately 1 trillion word tokens of text from publicly accessible Web pages."

Internet Archive
: Also known as the "Wayback Machine". Probably the largest repository of web-crawl data in existence. Not directly packaged for use as language data.

Common Crawl
: Eight years of roughly monthly Web crawls, now averaging around 3 x 10^9 (mostly HTML) pages per month, assembled in archive-format files each containing around 50,000 pages or page metadata or page text

And of course you can do your own crawl

- And there are lots of tools available to help

14. Open questions, changing times

The Web has made the impact of copyright in the digital domain even harder to figure out

- The copy your browser makes when you read a page is probably covered by Fair Use
- But what if you download it so you can read it offline?

Is a language model a derived work?

- The UK recently (2014, in the **Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014** act) allowed copying and processing of copyrighted material "for text and data analysis" as long as the sources were acknowledged and the research involved was "non-commercial"
- But the general question has not been definitively answered

What is the legal status of `robots.txt`?

- This is a standardised means of signalling to web crawlers *not* to harvest a particular page or web-site

15. Copyright and (personal) ethics

The most emotionally-charged aspects of digital copyright relate to music and video

- And those questions are out of scope for this course

But copyright on text not only affects you as an NLP scientist

- in ways outside your control
- In terms of what is and is not available for you to work with

But also it already affects you

- in ways you *do* control

You can contribute to the Open Science movement

- By putting a Creative Commons license on everything you put on the Web
  - Just as I've done with these slides
- By publishing your work in Open Access outlets
  - and putting any datasets you create in open respositories
- By using appropriate persistent identifiers whereever possible
  - For yourself
  - Your publications
  - Your data