Socially impactful digital libraries and/as challenges to NLP application

Mark A. Matienzo / @anarchivist
Stanford University Libraries
nlp4arc, Chapel Hill, NC, 2 February 2018 / #nlp4arc
Understanding social impact

- “the net effect of an activity on a community and the well-being of individuals and families” (Centre for Social Impact)
- “a significant, positive change that addresses a pressing social challenge” (University of Michigan Ross Center for Social Impact)
Ethics, social impact, and NLP

- Small but growing area of research
- Hovy and Spruit 2016 provide pivotal framing
  - Impact of NLP on social justice
  - Language as proxy for human behavior
  - Negative impact factors: exclusion, overgeneralization, bias confirmation, topic overexposure, and dual use
Social impact and digital libraries

- DLs as platforms for humanitarian information (Witten 2005)
- Frameworks for social impact of DLs: Tanner and Deegan 2011, Calhoun 2014
- **Socially impactful digital libraries** have a focus to improve society
- Emphasizes libraries as non-neutral actors (cf. Bourg 2015)

Calhoun’s framework of DL social roles
Socially impactful digital libraries at Stanford

- - -

- Portfolio development as a strategic goal; occasionally referred to as “humanitarian data”
- Broader context of similar projects with loose connection both within Stanford Libraries and across the university
- Two specific projects for this presentation
  - Digital Library of the Middle East
  - Virtual Tribunals
Digital Library of the Middle East

---

https://dlme.clir.org/

- Partnership between the Council on Library and Information Resources and the Antiquities Coalition, with Stanford Libraries as technical partner
- Response to conflict and loss of life in Middle East and increased looting of and trafficking in cultural heritage materials
- Broad and expansive vision outlined to stakeholders and funders
- Stanford component focuses on providing discovery platform for cultural heritage material from the Middle East and North Africa
DLME prototype and data challenges

- Aggregated metadata from 7 institutions (135K+ items)
- Common entities but sparse records, differing practices
- Variants in vocabulary, transliteration, language

https://spotlight.dlme.clir.org/
Virtual Tribunals

Collaboration between Stanford Libraries and the WSD Handa Center for Human Rights and International Justice

Goal is to build a comprehensive digital library platform for materials from international criminal tribunals and truth and reconciliation commissions

Current project focusing on materials from East Timor tribunal in four languages (English, Portuguese, Indonesian, Tetun)

Planned expansion into other tribunals (means other languages)

Entities include people, places, types of crimes/conflict, and many cross-references and citations

http://stanford.io/2DVwtAh
Challenges: underresourced languages

- Underresourced languages in an NLP sense are those which are missing machine-readable language data.
- Not all underresourced languages are “endangered”.
  - Consider Arabic (El-Haj 2015) and Sorani Kurdish (Walther and Sagot 2010).
- “All ASEAN languages are underresourced” (Ye 2016).
- Hovy and Spruit 2016 acknowledge commercial incentive to focus on overexposed languages (e.g. English).
Challenges: consistency

- Variation in transliteration schemes, orthography, etc. across communities represented in data
- Use of LC transliteration in subject headings, place names, etc. may not reflect more familiar or common transliteration
Challenges: entity extraction, reconciliation, etc.

- Particularly a challenge for underresourced languages
- Specialized concepts (e.g. projects like Virtual Tribunals)
- Impact of concept reconciliation as vector for introduction of bias or oppressive language
Challenges: avoiding reproduction of bias

- - -

• Hovy and Spruit 2016 identify introduction of bias as a major factor for production of linguistic data sets

• With regard to specific projects mentioned, major concern is how to avoid or eliminate Eurocentric and imperialistic biases

• Thought experiment: What is the social impact of using the Bible to produce corpora for a less-resourced language? ...especially if that community was impacted by Christian imperialism?
Challenges: ethical tool selection

- Promising tools exist that have been developed or funded in service of military and intelligence needs
- Anderson et al. 2012 identifies DARPA workshops as critical to development of computational linguistics research
- Hovy and Spruit 2016 consider government and military involvement in computational linguistics be worth a closer study
- How do we help LIS/archives practitioners to make informed choices?


• Tanner, Simon, and Marilyn Deegan. 2010. Inspiring Research, Inspiring Scholarship. JISC.


• Ye Kyaw Thu. 2016. “Challenges of Natural Language Processing Research for Under-Resourced Languages, using Myanmar Language as an Example.” Conf. on Khmer NLP, Phnom Penh.
Thank you!

Mark A. Matienzo / @anarchivist
Stanford Libraries
nlp4arc, 2 February 2018