Hackathons, Shared Tasks, and Papers

Collaboration and Interdisciplinarity with Data on Complex Problems

Myrthe Reuver

Image: Saydung89

First: Who am I?

PhD Candidate @ CLTL, in computational linguistics or Natural Language Processing (NLP). Supervisors: Antske Fokkens (CLTL @ VU), Suzan Verberne (LIACS @ Leiden).

- NLP is “Teaching computers how to deal with language”, related to Computer Science & AI.
- Spell checking, web searches, auto-fill… all use NLP.
- Language, and working with language data, is very complex.
  → nuances in meaning, pragmatics, normalization..
- You can find me on Twitter (@myrthereuver), where I often tweet about my work and field.

(Warning: these are only my personal experiences, not meant as universal representation of my field or of these events. Any errors or misrepresentations are my own.)

Second: What do I do?

(Viewpoint) Diversity in News Recommendation

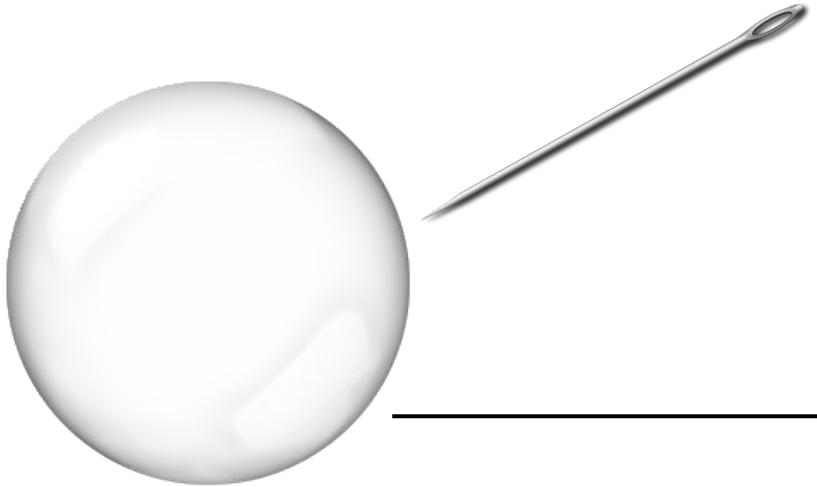
- Popping “filter bubbles”:
- Democratic debate online is becoming difficult due to online news consumption in **recommender systems** recommending only similar viewpoints and ideas (see also: “fabeltjesfuik” van Zondag met Lubach).

Interdisciplinary: co-PhD Nicolas Mattis works in social science, postdoc Marijn Sax works on ethics.

My job: automatically detecting, and diversifying recommendations, of viewpoints and ideas in news texts.

Third: What will I talk about?

The (a)typical approaches in NLP concerning data, and how they (do not) work for my project, science, and (interdisciplinary) collaboration.



NLP: a young, dynamic, and strange field

- I work in the VU Humanities department, but my work is related to Computer Science for many reasons.
  - working with code, math, and sometimes more of an engineering approach ("*how can we build a system that does X*?") rather than hypothesis-driven research.

- Quite a young field (1950s), and explosive growth the past few years (2010s).

- Recent community discussions within NLP:
  - Ethics, involvement of large private tech companies.
  - Do we want "slow science" rather than "flag-planting"?
  - **NLP is not very good at carefully assessing and curating data, leading to problems especially in very large pre-trained models, trained on millions of texts. (Paullada 2021, Bender 2020)**

Paullada et. al. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research / Bender et. al. 2020. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? "

United in Approach, Divided by Tasks

- NLP has many collaborative forms of working with data:
  - **(Shared) Tasks** → different groups perform one "task" with the same data, and compare their results during a workshop or conference. Which approach works for the task (e.g. detecting claims, or summarizing texts)? There is one "test set", an independent test for the models.
  - **Hackathons** → different groups compete on "solving" one problem or question, with a jury picking a winner.
  - **Benchmarks** → datasets, often stored centrally, used to "test" whether model Y performs better than model X on a task, like "claim detection". Different researchers can submit a "solution" any time they see fit, and see how their solution scored compared to other researchers in a "leaderboard".

- But.. these also have inherent **competition** rather than collaboration.
  - "winning" teams, "best" score on a benchmark is "State of the Art" (SOTA) model, etc.
  - recently, people have been noticing this & its downsides

see: the EMNLP 2020 Negative Results workshop had a panel on leaderboard-ism (and video).

Example: Metaphor Detection

Metaphor Detection

Metaphors is a form of figurative language used pervasively in our everyday lives. Consider the following examples:

... it would change the **trajectory** of your legal career ...

... Washington and the media just **explodes** on you, you just don't know where you are at the moment ...

... those statements are **deeply** concerning ....

... Out of the abundance of the heart, the mouth speaks, and the hand **tweets** ...

... the fake news were personalized and delivered above the **radar** and beyond the **radar** ...

The goal in this shared task is to detect, at word level, all content-word metaphors in a given text (we will also have a separate evaluation for just the verbs, as many researchers are working specifically on verb metaphors). We will use two datasets: (1) a subset of ETS Corpus of Non-Native Written English, which contains essays written by test-takers for the TOEFL test and was annotated for argumentation relevant metaphors in [Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2018. A Corpus of Non-Native Written English Annotated for Metaphor. NAACL]; (2) VU Amsterdam Metaphor Corpus (VUA) dataset (also used in the 2018 shared task), which consists of text fragments sampled across four genres from the British National Corpus (BNC): Academic, News, Conversation, and Fiction. The data is annotated according to the MIPVU procedure as described in [Gerard Steen, Aletta Dorst, Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. A Method for Linguistic Metaphor Identification. Amsterdam: John Benjamins.]

Important Dates

- January 12, 2020: 1st CFP for the shared task; CodaLab competition is open; training data and auxiliary scripts can be downloaded
- February ~~12~~ 14, 2020: 2nd CFP for the shared task; test data can be downloaded and results submitted; performance will be tracked on CodaLab dashboard
- ~~March 22~~ ~~April 11~~ April 16 11:59pm, 2020: Last day for submitting predictions on test data
- ~~April 18~~ April 23, 2020: Papers describing the systems are due for paper submission information

Does this work?

- Competition can be nice:
  - Incentive to make a “better” (better performing) model;
  - large strides have been made recently (rise of LLMs), and all in a short period of time;
  - Easy uniform comparison of models and approaches on benchmarks.

- But “leaderboard-ism” can also lead to:
  - Reward of only one type of research (“SOTA-chasing”) and not careful reflection on what, or why, we are measuring and building, or other aspects of models, from environmental impact to generalizability (Mihalcea, 2020).
  - The leaderboard becomes the means to an end, rather than a way to measure “task proficiency”.

Rada Mihalcea. 2020 “The ups and downs of word embeddings”. Talk at Negative Results Workshop EMNLP.

Diversity in news recommendation

I personally noticed “task-ification” is **also** not useful when dealing with complex societal problems. Why?

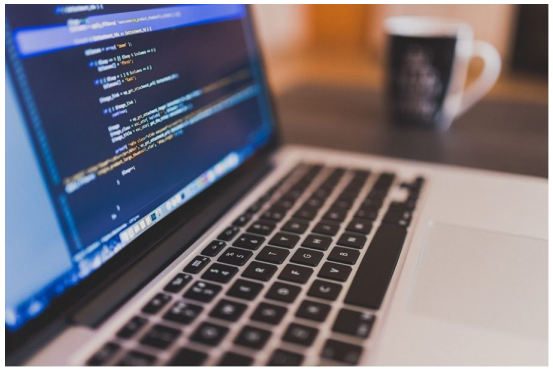
- **fragmentation of literature and ideas:** for “viewpoints”: a large set of tasks is relevant, with each their own definitions, datasets, and benchmarks:
  - *stance detection, argument mining, “perspectives”*
- **definitions:** Often, these tasks are not connected to theory or ideas from psychology, communication science, etc., but rather aimed at what is easy to measure or capture.
- **evaluation** across different tasks is difficult: different metrics (F1? precision?), benchmarks, etc.

Reuver et. al. (2021) “No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems” Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation.

My (interdisciplinary) experiences

1. **hackathon 1 - diverse recommendation (24 hours!)**

- NPO had a task: diversity in recommendation on NPO Start
- 5 junior teams, the goal: making a prototype of a diverse news recommender, with data provided.
- Our team had a Computer Scientist, Philosopher, Social Scientist, and Computational Linguist.
- We won! We wrote a blog about it,
  - working with different disciplines is difficult:
  - we remained separate islands, working on sub-parts
  - (e.g. data analysis, definitions and terminology, etc.).
  - Integration of these different ideas is hard.



2. Hackathon 2: EACL “Hackashop”

- a-typical hackathon: 3-week long, remote work on different challenges related to news.
- No judges: rather, every team wrote a short proceedings paper.

- Our work (mine & co-PhD Nick, social scientist): **Implementing metrics based on democratic theory in a news comment recommender.**
- Here, collaboration went better: theory, code, and ideas all came together.
- Time that collaboration took & theory-driven work meant, however, we had a very small results section.

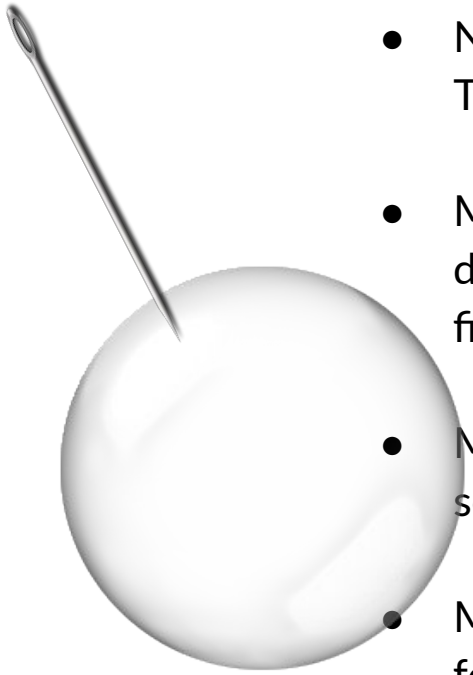
Myrthe Reuver & Nicolas Mattis. “Implementing Evaluation Metrics Based on Theories of Democracy inNews Comment Recommendation(Hackathon Report).” Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation.

3. “Paper in a Day”

- Project Group of the RecSys project wanted to work on a shared paper
- We followed a new approach: “Paper in a Day”, to write it in one day:
  - leaving a whole day for writing & talking;
  - setting up a structure of points we want to write about;
  - discussing and “walking around” virtually in a virtual room (gather.town).

- It was a really nice “getting to know each other” exercise as well

Conclusion



- NLP has an a-typical approach to data usage in science: Shared Tasks, hackathons, and benchmarks.
- Many of these are aimed at collaboration, but also competition, and discussion has come up about how useful this is for science & the field.
- My interdisciplinary collaboration has used such approaches: sometimes with success, sometimes not.
- My take-away: reflect on whether “best practices” of your field work for your research; collaborate, especially across fields and when concerning real-life problems with complicated theory.

Thank you!

I hope you found my talk fun and/or useful.

:-)

Any questions?