

From data collection to output evaluation:

the challenging journey towards fair, robust and

interpretable NLP systems

Els Lefever 27 Nov 2024





DEPARTMENT OF TRANSLATION, INTERPRETING AND COMMUNICATION



LT3 TEAM

8 professors4 postdocs



www.lt3.ugent.be





16 predocs2 IT support

GHENT CENTER FOR DIGITAL HUMANITIES



Cinema Ecosystem (CINECOS) More information







Computational Literary Studies Infrastructure (CLS INFRA) More information



CUNE-IIIF-ORM: Towards an Internationally Image Interoperable **Corpus of Cuneiform** Tablets

More information



DARIAH in Belgium (DARIAH-BE) More information





More information



Database of Byzantine Book Epigrams (DBBE) More information



Diplomata Belgica More information



Digital Literacy in the Faculty of Arts and Philosophy

More information



Everyday Writing in Graeco-Roman and Late Antique Egypt. A Socio-Semiotic Study of Communicative Variation (EVWRIT)



RESEARCH LINES

Language technology

coreference resolution, cross-lingual transfer models, detection of events, sentiment, irony, arguments and emotion in (financial) news data and social media

Translation technology

machine translation, post-editing, human-computer interaction, translation quality, translation difficulty assessment and gender-inclusive translation

Digital Humanities

digital text analysis tools for research in the humanities and social sciences

Language and translation technology for educational applications automatic writing evaluation, readability assessment, vocabulary and example selection

automatic writing evaluation, readability assessment, vocabula for SLA, MT for language learning

Terminology



automatic (multilingual) terminology extraction and terminology management

<u>OUTLINE</u>



Interpretable









NLP and Machine learning





AI / Language and Translation Technology







Sentiment Analysis





HOW DOES A **COMPUTER LEARN LANGUAGE?**

- 1. Linguistic rules created by experts: rule**based** and lexicon-based approaches
- 2. Rules are learnt based on examples: data-based approaches

= Machine learning: ""giving computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959).





Machine learning



Recent ML: Neural systems

Training data



Neural network









Prediction



$\label{eq:FR} \begin{array}{l} FR \rightarrow NL \\ un \ navire \rightarrow een \ schip \end{array}$



Step 1: train (large) language model

> Trained to predict the statistically most probable next word (on a massive text) corpus)

Whisk together the flour, baking soda, and a pinch of [???] in a large bowl.



How does it work?

- Computers cannot work with text > we represent words as numeric vectors \triangleright computers can work with
- Those numbers contain information about the meaning of words, deduced from the contexts in which these words occur (in massive text collections)





Language model

How does it work?

Step 2: Fine-tune large language model



Fine-tuning

train large language
model for specific task
based on manually
labeled data (e.g., for
sentiment analysis)





Fair NLP systems





Bias in NLP





BIAS in NLP systems



UNIVERSITY



Model generates output for new data

Selection and model bias

The New York Times

Opinion

OPINION

Artificial Intelligence's White Guy Problem

By Kate Crawford

June 25, 2016



English

"The doctor asked the nurse to help <u>her</u> with the procedure"





Spanish

"<u>El</u> doctor le pidió a la enfermera que le ayudara con el procedimiento"





Currey & Hsu, EMNLP 2022 https://www.amazon.science/blog/datasethelps-evaluate-gender-bias-in-machinetranslation-models

Selection and model bias

Complete this sentence: the man worked for a long time as a ...

\$ The man worked for a long time as a doctor dedicating his life to helping others and improving their health.







Currey & Hsu, EMNLP 2022

Janiça Hackenbuchner



DeBiasByUs



Joke Daems









Alessandra Teresa Cignarella



RAINBOW



Els Lefever



RESEARCHING STEREOTYPES TOWARDS _GBTQIA+ PEOPLE WITH MULTILINGUAL NLP

Why stereotypes?

1. They are at the base of the "Pyramid of Hate"

2. They can help in preventing Hate Speech in its early-manifesting phases

Why LGBTQIA+?

They are among the most critically targeted groups online 1.



2. Only a few studies so far (most work explores stereotypes about) ethnicity, gender or religion...)

Genocide

The act or intent to deliberately and tematically annihilate an entire people

Bias Motivated Violence

Murder, Rape, Assault, Arson, Terrorism, Vandalism, Desecration, Threats

Discrimination

Economic discrimination, Political discrimination, ucational discrimination, Employment discrimination, Housing discrimination & segregation, **Criminal justice disparities**

Acts of Bias

Bullying, Ridicule, Name-calling, Slurs/Epithets, Social Avoidance, De-humanization, Biased/Belittling jokes

Biased Attitudes

Stereotyping, Insensitive Remarks, Fear of Differences, Non-inclusive Language, Microaggressions, Justifying biases by seeking out like-minded people, Accepting negative or misinformation/screening out positive information

RESEARCH QUESTIONS & OBJECTIVES

- **1.** How do we generalize and define stereotypes (towards) LGBTQIA+ people)?
- 2. How do we implement fairer and more inclusive AI systems?
- 3. How can we use NLP applications to foster positive online behavior in younger generations with regards to LGBTQIA+ individuals?





Ecological footprint





Ecological footprint of large language models



De Standaard, 15/03/2023

- GPT-3: trained on hundreds of millions of text pages (45 terabyte text);
- 175 billion of parameters from data; flights from London to New York
- While training, the algorithm deduces - Training phase corresponds to 600





Fair and robust NLP

systems





language and translation technology team



English vs low-resourced languages

=> State-of-the-art NLP models are English-centric



GPT3: 92,1 = English



GPT3: 92,1% training data





Els Lefever



LANGUAGE MODELS FOR LOW-

RESOURCED LANGUAGES



Orphée De Clercq **Aaron Maladry**





29

Bridging the language gap for NLP

- . Investigate approaches to improve language models for low(er)-resourced languages
 - > +7000 languages, only a few dozens profit from research in NLP (Joshi et al. 2020)
 - > research language models for low-resourced languages:
 - using pretrained multilingual language models
 - adapting them
 - training new models from scratch



Cross-lingual transfer



Idea: use (labeled) data from one (or more) source languages to solve a problem for a low(er)-resourced target language



Evaluation for this task on target language

EXALT: multilingual data set for emotion

- Explainability for cross-lingual emotion in tweets
- Trained on English emotion data
- Evaluated on a wide range of target languages (ao Dutch, Russian, Spanish, French, Japanese, ...)

	Trigg
Emotion label	Trio
Anger ^[1] Sadness ^[2] Fear ^[3] Joy ^[4] Love ^[5] Neutral ^[6]	
Discard ^[7]	Sta









words

ay from me and mines or u gonna get hurt.

NLP FOR ANCIENT LANGUAGES



33

* isuBisi: + apeicTex "NHE TAE DERATAE "H" TPACE DERATO THE TENAY L'AN : + אישי איז גישא גיוריב שע טע דייי וישי וישי אי אי אי אי אי אי אי אי גע געיג אייי אי גע געייאיי + o Datathiar Bu Buton Topickars Francis Ballo To Ballo The Ano Bropiar +

in a marthing & a many man and provider a standar mare in any a side topo and fore

the institut in the new man a bus the part of the charge 30 of the promotion have not use was xand a set the part of the past of state of a set of the and the man prior inguites , as the the to the to the to the bear of the part of the part of the part of the total Preserve on the for in iter day an in mapling is to ante wout bu the dry by igna paparili Hap comment in our owned the man maple riger with your of us in a generation of more your why as we as " top to see we and to go and " I to mar all the goat to les - Barrow marity provide my greath, Kargon his and in min kin mint in mich by the sais Tingue plapanga

CALL ON LABORATION AND NI BIRSON. pill lever Tourter FAR NUM, apry NORE YOONE a ster all deliver ig agritup + 2 ternet Smith us mar prover TOSTANO ROONATI 60 M SANE Tingoro A aron a NTTUT quely t

5 hubrary of the opin requirection life au man you out -Toior Jugar TENNY LANCE BADI TON TOAT 644 hrows is To randow areas and for capeidhor the war share to cherenor out דר אי איל ועודי ומולג אול ד מים ap 29000 person to the off of 11, is mine Lephque or alastipate of New wort leafar. (m. for the is wain where une on have main phycies Sartopo Figure por interior man dans סדראסד avy, himer queeres barres les narth Su. Mage maddyed 5 + +10 min Hode subplin to TPI repus de partespart norn Time " nove it typerates to ann and he lighty in mounding and & and TERIORH, SALA TEL THE TEL ST AN A THE THE STORE

do The rause ourporpeta -

ap rateine no yous o topar upa spoin former. and the routioc udaren to philubrauphi. he by ubournetice , good but up forto wot. harby thing toc KA 101 as to ano piadas. is tragismon roip tranuphar by ran Sproup. hyopohimiorporkailing the xupien. 64 Awarding the us are deepon to op the again inotopocouna blaans centide unixolau Laporni din

> e' pap sé dé mie real que kei O ais sun de dou ais tois pas reis E'we i had an E mal. & douge ai rois Q anas É Di Gty ... Va aura Obas main's andi אום שמי יווד יי עוד ידי שמדי לי עוד אין הי אוד ידי שמדי ד שנט וערע לאוצולי לאיר איד לעון

· epixudeadaut LEVOI ,0

nov Tal Dove x Topai noodafiero : +

oiapuso, diose Deos Gaorx Fos.

TO OT SO

crance

REILITER Platt

Hiel Eneno Harte

Howmanite

adipour tai gat thug

: rand ad apropries dup acon watch bk mop an opti-

michira

16 Hear

wip ford

Tigo ble

6006000

in provident

xprogor

pi toi

4 ubrobacce

Colin Swaelens

Pranaydeep Singh

TESSERACT

Tel Ter The The Way of X Ter with on Print Way of the

- Interdisciplinary project: "Interconnected texts": a graph-based computational approach to metrical paratexts in Greek manuscripts (NLP (LT3), Greek literature, Greek linguistics, computer science)
 - NLP: Measuring Orthographic and Semantic Similarity between Byzantine Greek Epigrams





DATA USED TO TRAIN DBBERT





Fine-tuning part-ofspeech + fine-grained morphology

- Morpheus: rule-based system
- RNN Tagger: best- performing • for AG
- Freq. baseline: most occurring label / token
- <u>DBBErt</u>: fine-tuned embedding of our DBBErt













CUNE-IIIF-ORM



Katrien De Graef

Els Lefever







Towards an Internationally Interoperable Corpus of Cuneiform Tablets

- IIIF an image and text API
- OCR automatically reading cuneiform texts
- NLP annotate and analyze Akkadian texts (Ghent University) > Fully annotated Old Babylonian (c. 2000-1600 B.C.E.) Akkadian letters





AKKADIAN

- East Semitic language
- Written with the cuneiform script
- In use for more than 2500 years
- Dominated modern-day Iraq





NLP FOR CUNEIFORM AKKADIAN

ML experiments for Part-of-Speech tagging and morphological annotation:

- Embedding models: **Multilingual BERT** Semitic PLM: Arabic, Hebrew Japanese

- Avg. accuracy results (5-fold on 10K tokens)
 - PoS (transliterated) Arabic: 94,1 % Japanese: 93,4 % mBERT: 90,3%





PoS + morphological tags Multilingual: 71,0 % Arabic: 76,2 %

NLP FOR CUNEIFORM AKKADIAN

Problems

- Low-resourced language
- Few machine readable texts
- Inconsistent formatting and missing annotation standard

Solutions

- > Specialists gathering data
- > Develop UD standards

>> Further investigate impact of:

- different language models / combinations of languages
- adding similar data to train a first Akkadian language model





> Support with larger Semitic languages (Arabic and Hebrew)



Interpretable NLP systems





language and translation technology team



Data

Output

Aaron Maladry

Els Lefever

IRONY DETECTION

Cynthia Van Hee

Véronique Hoste

46

Irony detection

- Manual annotations by trained linguists
- Task: which tweets are ironic and how is the irony realised?

literal sentiment: positive ("can't wait")

intended sentiment: negative ("go to the dentist")

Target Megative

go to the dentist tomorrow! #sarcasm

Irony Detection: trigger words

- trigger word annotation
 - By humans
 - By systems
- advantages:
 - align with system
 interpretability

Irony detection: explanations by humans and machines

What do trigger words mean? Why these words? => open to interpretation

- Generate & evaluate explanations
- Compare human and generated explanations

Ironic tweet: Loooovvveeeeeee when my phone gets wiped

Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.

Background knowledge:

- When a phone gets wiped, all personal data and information is lost.
- People do not like losing access to their personal data on their phone

Irony detection: explanations by humans and machines

Evaluate? Explanation ranking by other group of humans

- > works very well for English
- > GPT models ranked higher than humans
- > other fine-tuned generative explanations are indistinguishable from human explanations > Next: Dutch explanations !?

Lot of ongoing research and remaining challenges to investigate more fair, robust and interpretable NLP systems:

 carefully curated data sets covering different languages, minority groups, domains, text genres and language variants (historical, dialects, ...)

cross-disciplinary research

Els.lefever@ugent.be

https://lt3.ugent.be/

