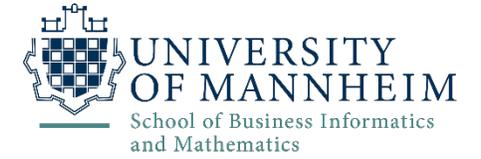


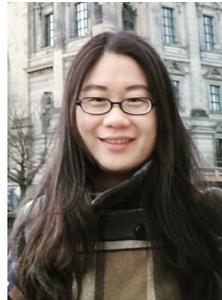
Reproducibility in Natural Language Processing: what can we learn from NLP?

Simone Paolo Ponzetto



Hi there!

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**



Hi there!

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**
- We are part of the larger **Data and Web Science** fleet @ Uni Mannheim

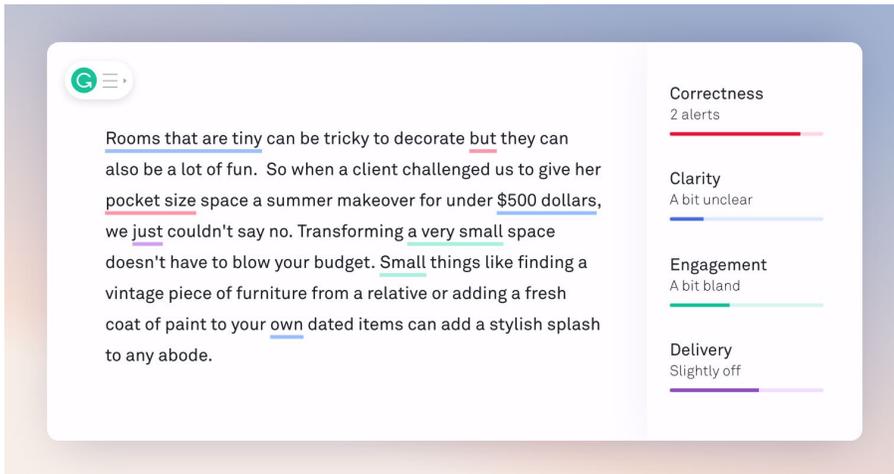


Today on the menu!

- Natural Language Processing?
- Reproducibility in NLP
 - Community-wide evaluation exercises (a.k.a. “shared tasks”)
 - Open publication models
 - Open data & software
- What can we learn from this?

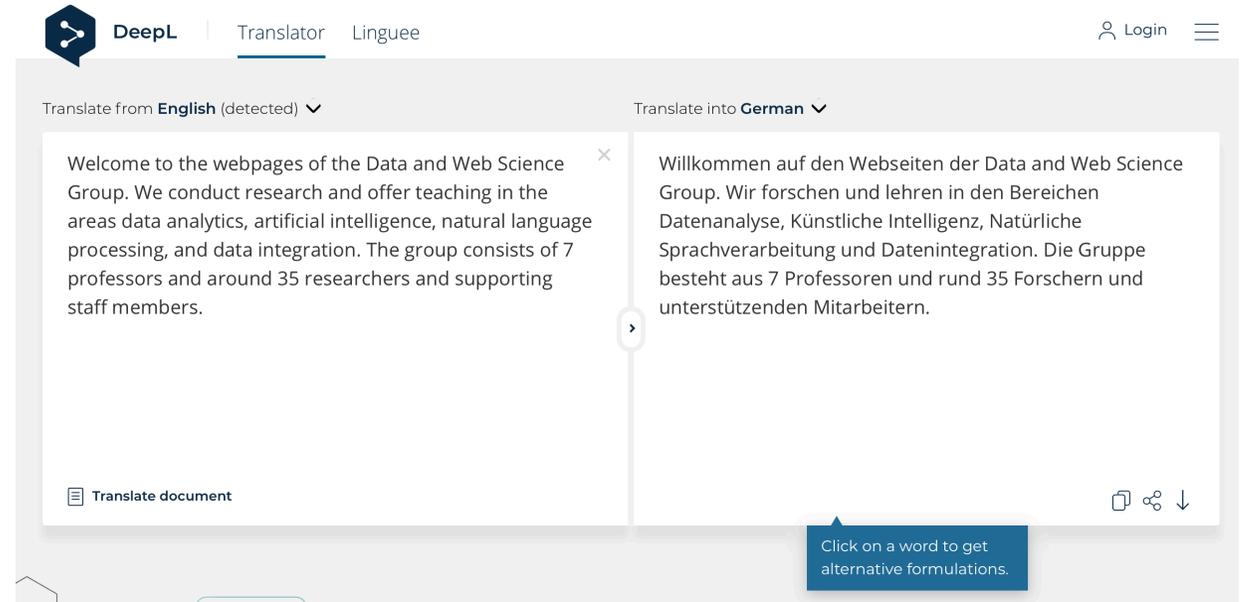
Natural Language Processing: some initial thoughts...

- Methods to automatically process (i.e., understand and generate) natural language data



Rooms that are tiny can be tricky to decorate but they can also be a lot of fun. So when a client challenged us to give her pocket size space a summer makeover for under \$500 dollars, we just couldn't say no. Transforming a very small space doesn't have to blow your budget. Small things like finding a vintage piece of furniture from a relative or adding a fresh coat of paint to your own dated items can add a stylish splash to any abode.

Correctness	2 alerts
Clarity	A bit unclear
Engagement	A bit bland
Delivery	Slightly off



DeepL | Translator | Linguee

Translate from English (detected) | Translate into German

Welcome to the webpages of the Data and Web Science Group. We conduct research and offer teaching in the areas data analytics, artificial intelligence, natural language processing, and data integration. The group consists of 7 professors and around 35 researchers and supporting staff members.

Willkommen auf den Webseiten der Data and Web Science Group. Wir forschen und lehren in den Bereichen Datenanalyse, Künstliche Intelligenz, Natürliche Sprachverarbeitung und Datenintegration. Die Gruppe besteht aus 7 Professoren und rund 35 Forschern und unterstützenden Mitarbeitern.

Translate document

Click on a word to get alternative formulations.

A few applications (from our everyday's life!)

Spelling
correction

Text completion

Grammar
checking

Speech-to-text
and vice versa

Dialogue systems

Question
Answering

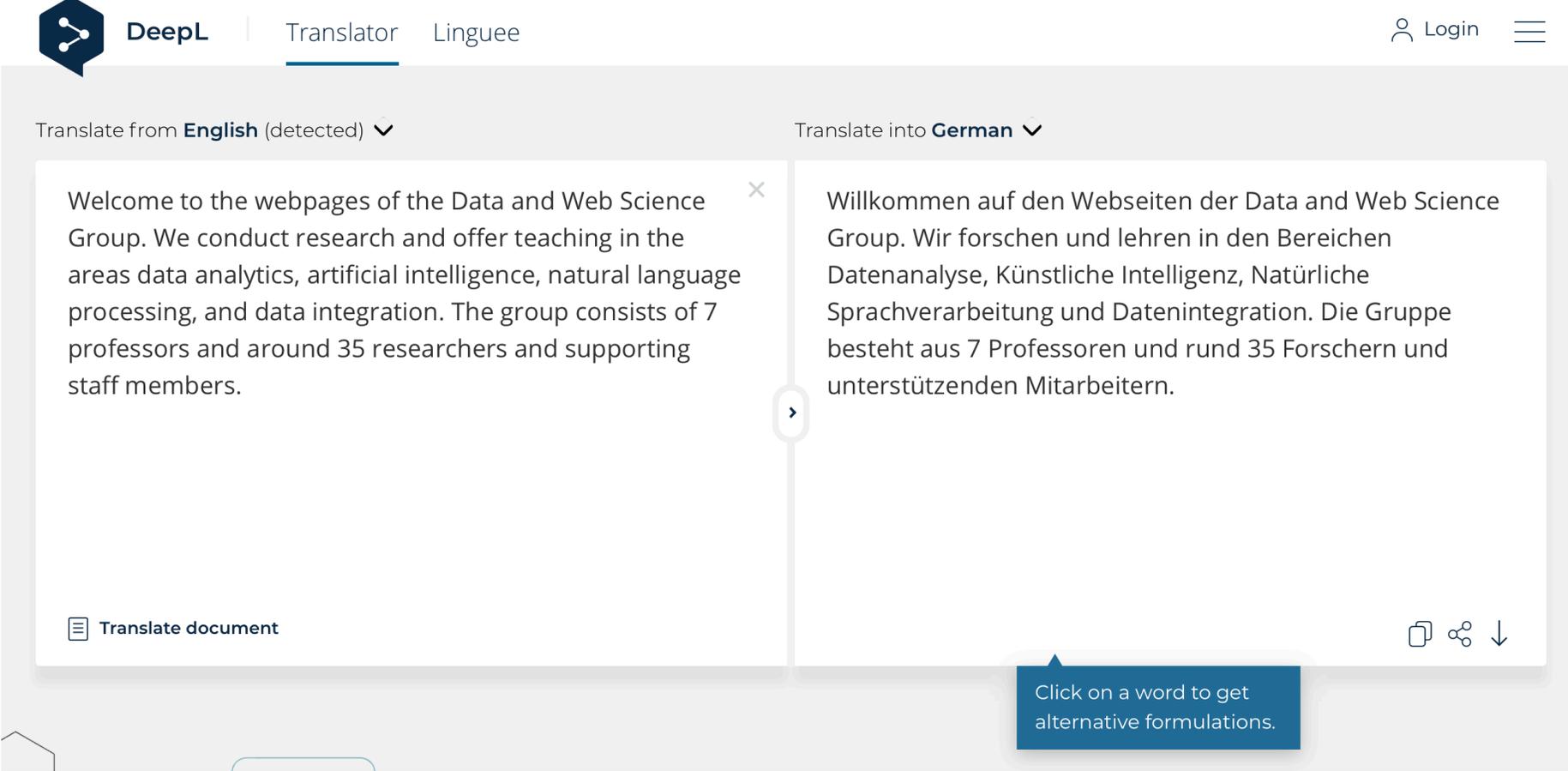
Summarization

Machine
translation

Evaluation of NLP systems

- The goal of NLP evaluation is to measure one or more *aspects* of an algorithm or a system. Examples:
 - **Performance** (how *good* it is)
 - **Efficiency** (e.g., speed, resource requirements)
- To **establish performance** and **comparability of different systems** we need to define a *experimental setting*
 - Evaluation metrics
 - Evaluation data (language data resources)

Example: machine translation



The screenshot shows the DeepL Translator interface. At the top left is the DeepL logo, followed by the text "Translator" and "Linguee". On the top right, there is a "Login" button and a menu icon. The main area is divided into two columns. The left column is labeled "Translate from English (detected)" and contains the text: "Welcome to the webpages of the Data and Web Science Group. We conduct research and offer teaching in the areas data analytics, artificial intelligence, natural language processing, and data integration. The group consists of 7 professors and around 35 researchers and supporting staff members." Below this text is a "Translate document" button. The right column is labeled "Translate into German" and contains the German translation: "Willkommen auf den Webseiten der Data and Web Science Group. Wir forschen und lehren in den Bereichen Datenanalyse, Künstliche Intelligenz, Natürliche Sprachverarbeitung und Datenintegration. Die Gruppe besteht aus 7 Professoren und rund 35 Forschern und unterstützenden Mitarbeitern." At the bottom right of the interface, there are icons for document, share, and download. A blue callout box at the bottom right contains the text: "Click on a word to get alternative formulations."

Evaluating Machine Translation

English	German
Diverging opinions about planned tax reform	Unterschiedliche Meinungen über die geplanten Steuerreformen
The discussion around the envisaged major tax reform continues .	Die Diskussion um die vorgesehene grosse Steuerreform dauert an .
The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .	Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafür aus , wesentliche Teile der für 1999 geplanten Reform vorzuziehen .

Reference translation 1:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:

The American [?] international airport and its office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:

US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

Shared tasks?

- In a nutshell, a **competition-like, community-wide evaluation exercise**
- **Shared experimental setting**
 - Data
 - Metrics
- Open for virtually anyone to
 - **Participate**
 - **Describe approach** into a report (typically a workshop paper)
 - **Present** in front of a large audience of experts and fellow colleagues

Shared task in 2020s

- Data Science competitions are all the rage these days: see Kaggle
- NLP has a long tradition of organising competitions of this kind
- First attempts with MUC / TREC (through funding and support from NIST and DARPA)

Simone Ponzetto / Data Science in Action

18.11.2022

Source: Wikipedia



Text REtrieval Conference

...to encourage research in information retrieval from large text collections.

Abbreviation TREC

Discipline information retrieval

Publication details

Publisher NIST

History 1992; 30 years ago

Frequency annual

Website trec.nist.gov

Conference	Year	Text Source	
MUC-1	1987	Mil. reports	Fleel
MUC-2	1989	Mil. reports	Fleel
MUC-3	1991	News reports	Terr
MUC-4	1992	News reports	Terr
MUC-5	1993	News reports	Corp
MUC-6	1995	News reports	Neg Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

SemEval

[View on GitHub](#)



SemEval-2023

The 17th International Workshop on Semantic Evaluation

We are pleased to announce the following tasks for
SemEval-2023!

TASKS

Websites and contact information for individual tasks are given below.

Semantic Structure

- **Task 1: V-WSD: Visual Word Sense Disambiguation** ([contact organizers], [join task mailing list])
Alessandro Raganato, Iacer Calixto, Jose Camacho-Collados, Asahi Ushio, Mohammad Taher Pilehvar
- **Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER 2)** ([contact organizers], [join task mailing list])
Shervin Malmasi, Besnik Fetahu, Sudipta Kar

Discourse and Argumentation

- **Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup** ([contact organizers], [join task mailing list])
Giovanni Da San Martino, Jakub Piskorski, Nicolas Stefanovitch, Preslav Nakov
- **Task 4: ValueEval: Identification of Human Values behind Arguments** ([contact organizers], [join task mailing list])
Johannes Kiesel, Milad Alshomary, Henning Wachsmuth, Benno Stein
- **Task 5: Clickbait Spoiling** ([contact organizers], [join task mailing list])
Maik Fröbe, Tim Gollub, Matthias Hagen, Martin Potthast
- **Task 6: LegalEval: Understanding Legal Texts** ([contact organizers], [join task mailing list])
Prathamesh Ashok Kalamkar, Saurabh Kumar Karn, Sachin Malhan, Vivek Raghavan, Shouvik Kumar Guha, Ashutosh Modi

Medical Applications

- **Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data** ([contact organizers], [join task mailing list])
Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, Andre Freitas
- **Task 8: Causal medical claim identification and related PICO frame extraction from social media posts** ([contact organizers] [join task mailing list])
(temporarily anonymous)

Social Attitudes

- **Task 9: Multilingual Tweet Intimacy Analysis** ([contact organizers], [join task mailing list])
Jiaxin Pei, Francesco Barbieri, Vítor Silva, Maarten Bos, Yozen Liu, Leon David Jurgens
- **Task 10: Towards Explainable Detection of Online Sexism** ([contact organizers], [join task mailing list])
Hannah Rose Kirk, Wenjie Yin, Paul Röttger, Bertie Vidgen
- **Task 11: Learning with Disagreements (Le-Wi-Di), 2nd edition** ([contact organizers], [join task mailing list])
Elisa Leonardelli, Valerio Basile, Gavin Abercombie, Tommaso Fornaciari, Plank, Massimo Poesio, Verena Rieser, Alexandra Uma
- **Task 12: AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset** ([contact organizers], [join task mailing list])
Shamsuddeen Hassan Muhammad, Seid Muhie Yimam, Idris Abdulmumini, Said Ahmad, Saif Mohammad, David Ifeoluwa Adelani, Sebastian Ruder, Ousidhoum, Vukosi Marivate, Abinew Ali Ayele, Meriem Beloucif

CoNLL

- Conference on Natural Language Learning is a yearly meeting of Special Interest Group on Nature Language Learning (SIGNLL) of the Association for Computational Linguistics
- Since 1999, CoNLL has included a **shared task** in which *training and test data is provided by the organizers* which **allows participating systems to be evaluated and compared in a systematic way**

CoNLL: previous shared tasks

CoNLL

The SIGNLL Conference on Computational Natural Language Learning

Previous shared tasks

2019	Cross-Framework Meaning Representation Parsing	English	Proceedings
2018	Universal Morphological Reinflection	multilingual	Proceedings
2018	Multilingual Parsing from Raw Text to Universal Dependencies	multilingual	Proceedings
2017	Multilingual Parsing from Raw Text to Universal Dependencies	multilingual	Proceedings
2017	Universal Morphological Reinflection	multilingual	Proceedings
2016	Multilingual Shallow Discourse Parsing	English, Chinese	Proceedings
2015	Shallow Discourse Parsing	English	Proceedings
2014	Grammatical Error Correction	English	Proceedings

2013	Grammatical Error Correction	English	Proceedings
2012	Modelling Multilingual Unrestricted Coreference in OntoNotes	English, Chinese, Arabic	Proceedings
2011	Modelling Unrestricted Coreference in OntoNotes	English	Proceedings
2010	Hedge Detection	English	Proceedings
2009	Syntactic and Semantic Dependencies in Multiple Languages	multilingual	Proceedings
2008	Joint Parsing of Syntactic and Semantic Dependencies	English	Proceedings
2007	Dependency Parsing: Multilingual & Domain Adaptation	multilingual	
2006	Multi-Lingual Dependency Parsing	multilingual	
2005	Semantic Role Labeling	English	
2004	Semantic Role Labeling	English	
2003	Language-Independent Named Entity Recognition	English, German	
2002	Language-Independent Named Entity Recognition	Spanish, Dutch	
2001	Clause Identification	English	
2000	Chunking	English	
1999	NP Bracketing	English	

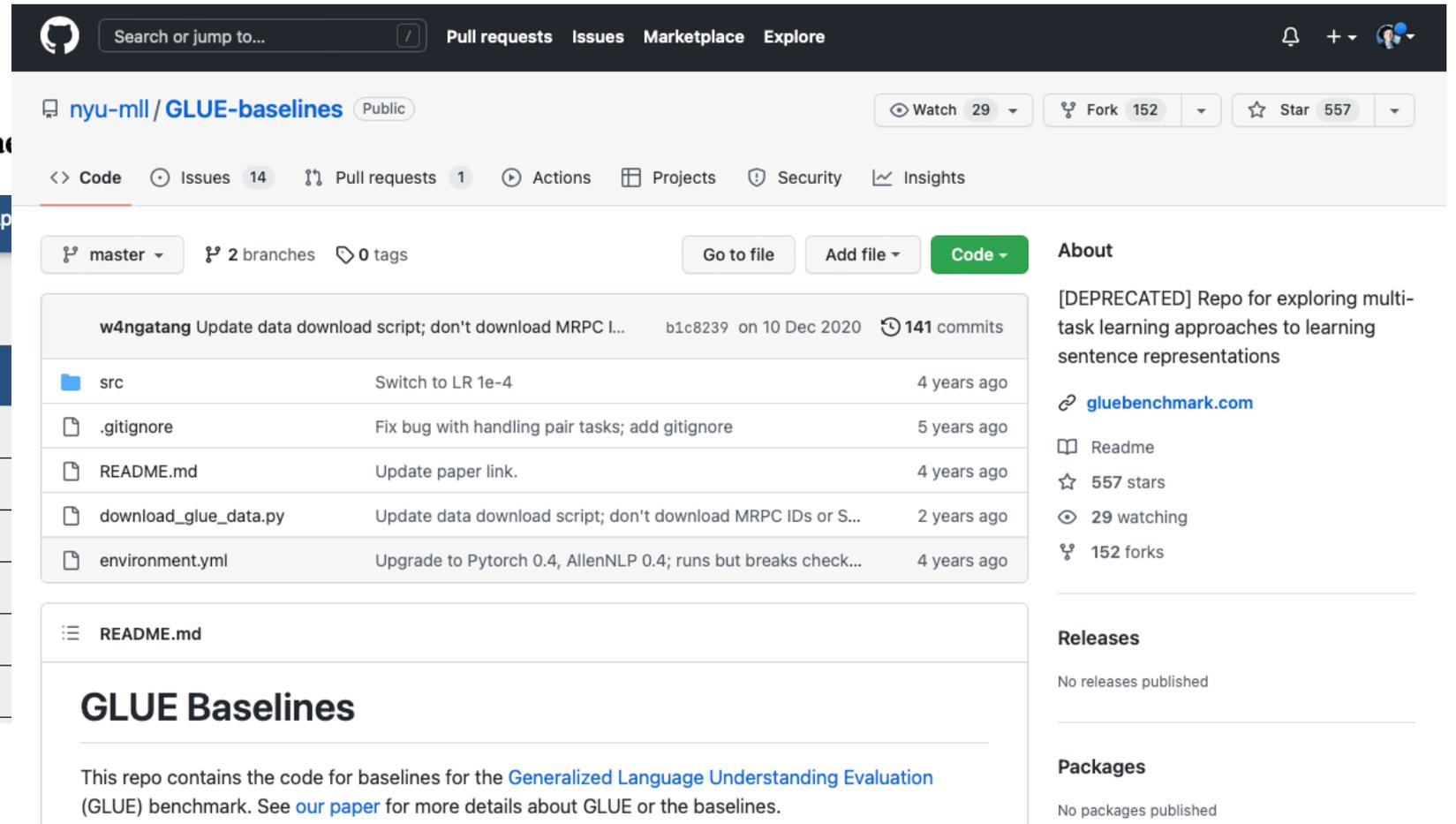
GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTAND- ING

Alex Wang¹, Amanpreet Singh¹, Julian Michael
Omer Levy² & Samuel R. Bowman¹

¹Cour
²Paul
³Deep
{ale
{jul
feli

Rank	Name	Model
1	JDExplore d-team	Vega v1
2	Microsoft Alexander v-team	Turing NLR v5
3	DIRL Team	DeBERTa + CLEVER
4	ERNIE Team - Baidu	ERNIE
5	AliceMind & DIRL	StructBERT + CLEVER
6	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4

Simone Ponzetto / Data Science in Action
18.11.2022



The screenshot shows the GitHub repository page for `nyu-ml/GLUE-baselines`. The repository is public and has 29 watchers, 152 forks, and 557 stars. It contains 14 issues, 1 pull request, and 141 commits. The repository is currently on the `master` branch with 2 other branches and 0 tags. The file list includes `src`, `.gitignore`, `README.md`, `download_glue_data.py`, and `environment.yml`. The `README.md` file is selected and shows the title `GLUE Baselines`. The description of the repository states: "This repo contains the code for baselines for the Generalized Language Understanding Evaluation (GLUE) benchmark. See our paper for more details about GLUE or the baselines."

<https://gluebenchmark.com> 15

ACL anthology



ACL Anthology

[FAQ](#) [Corrections](#) [Submissions](#)

Search...



Welcome to the ACL Anthology!

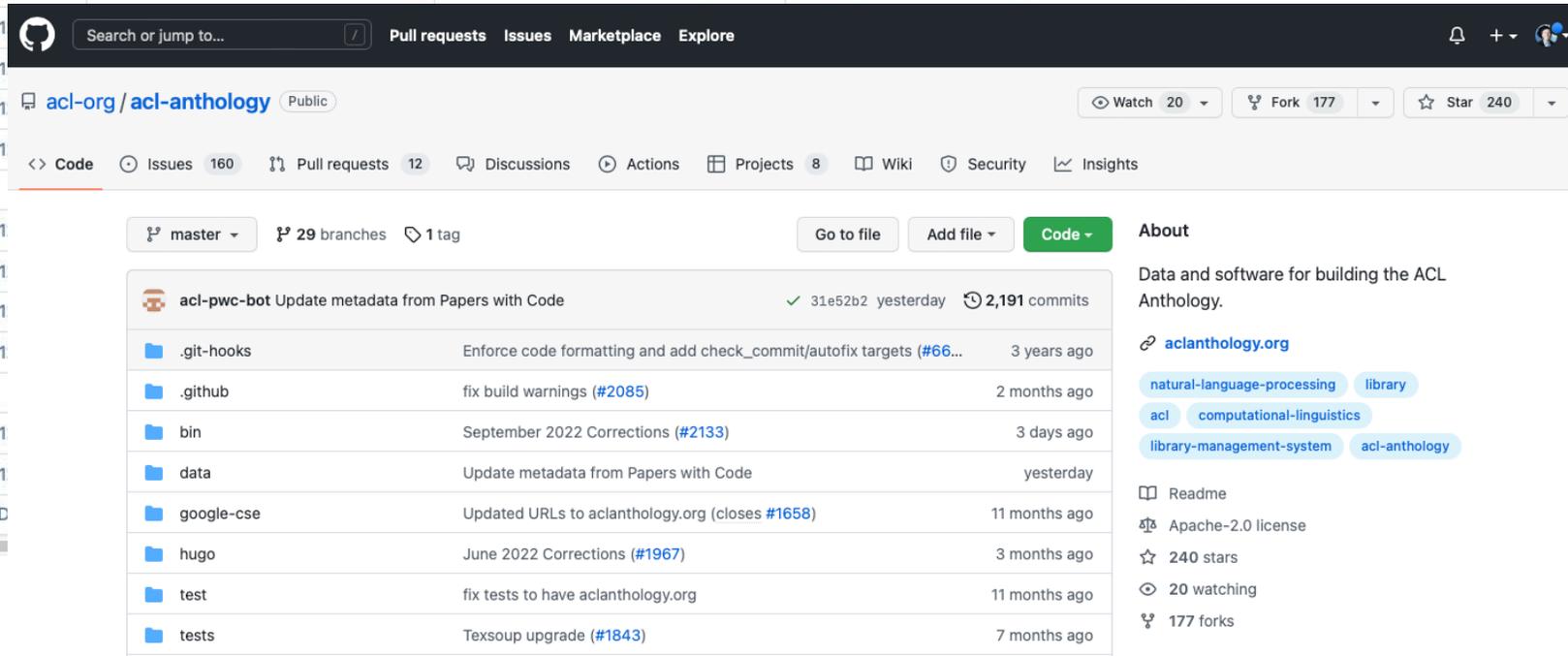
The ACL Anthology currently hosts 79296 papers on the study of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

Full Anthology as BibTeX (6.78 MB)

ACL Events

Venue	2022 – 2020	2019 – 2010										2009 – 2000										1999 – 1990										1989 and older									
AAACL	20																																								
ACL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
ANLP																						97 94 92	88	83																	
CL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
CoNLL	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
EACL	21	17 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	97 94 92	88	83																																			
EMNLP	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
Findings	22 21 20																																								
IWSLT	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
NAACL	22 21	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
SemEval	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
*SEM	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
TACL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
WMT	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
WS	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79																																				
SIGs											ANN BIOMED DAT DIAL ED																														



Simone Ponzetto / Data Science in Action

18.11.2022

<https://aclanthology.org>

Make the code available!



NLP & IR Group @ University of Mannheim
Natural Language Processing and Information Retrieval Group at the University of Mannheim
Mannheim, Germany <http://dws.informatik.uni-mannhei...>

Follow

Overview Repositories 30 Projects Packages Teams 2 People 14 Settings

Popular repositories

tg2019task Public archive TextGraphs-13 Shared Task on Multi-Hop Inference Explanation Regeneration Python 43 stars 6 forks	SemScale Public Forked from codogogo/topfish A tool for Semantic Scaling of Political Text (branch of Topfish, a suite of tools for Political Text Analysis) Python 17 stars 4 forks
RedditBias Public Forked from SoumyaBarikeri/RedditBias Code & Data for the paper "RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models" Python 8 stars 1 fork	crosstemporal_bias Public Forked from nfriedri/Investigating-Antisemitic-Bias-in-German-Parliamentary-Proceedings Detect racial biases within German parliamentary proceedings reaching from 1867 - 2020 Python 4 stars
redditbias_debias_conv_ai Public Forked from SoumyaBarikeri/debias_transformers	FairArgumentativeLM Public Code & Data for the paper "Fair and Argumentative Language



Search

Browse State-of-the-Art Datasets Methods More

Browse State-of-the-Art

9,360 benchmarks 3,684 tasks 80,023 papers with code

Computer Vision

Semantic Segmentation 171 benchmarks 3194 papers with code	Image Classification 376 benchmarks 2637 papers with code	Object Detection 250 benchmarks 2408 papers with code	Image Generation 197 benchmarks 1013 papers with code	Denoising 109 benchmarks 973 papers with code
---	--	--	--	--

See all 1347 tasks

Natural Language Processing

Language Modelling 373 benchmarks 2052 papers with code	Question Answering 156 benchmarks 1683 papers with code	Machine Translation 76 benchmarks 1635 papers with code	Sentiment Analysis 85 benchmarks 988 papers with code	Text Generation 233 benchmarks 851 papers with code
--	--	--	--	--

See all 637 tasks

How to promote this: make it appealing for researchers

Google Scholar



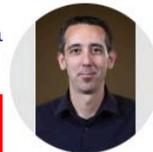
Samuel R. Bowman

Assistant Professor, [NYU](#)
Verified email at nyu.edu - [Homepage](#)
natural language processing representation learning compu
crowdsourcing AI alignment

TITLE

- GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding**
A Wang, A Singh, J Michael, F Hill, O Levy, SR Bowman
Proceedings of ICLR
- A large annotated corpus for learning natural language inference**
SR Bowman, G Angeli, C Potts, CD Manning
Proceedings of EMNLP
- A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference**
A Williams, N Nangia, SR Bowman
Proceedings of NAACL-HLT
- Generating sentences from a continuous space**
SR Bowman, L Vilnis, O Vinyals, AM Dai, R Jozefowicz, S Bengio
Proceedings of CoNLL
- SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems**
A Wang, Y Pruksachatkun, N Nangia, A Singh, J Michael, F Hill, O Levy, ...
Proceedings of NeurIPS

Google Scholar



Lluís Màrquez

Principal Applied Scientist
Verified email at amazon.com

Artificial Intelligence Natural Language Processing Machine Learning

FOLLOW

TITLE

- Introduction to the CoNLL-2005 shared task: Semantic role labeling**
X Carreras, L Màrquez
Proceedings of the Ninth Conference on Computational Natural Language ...
- Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling**
X Carreras, L Màrquez
Proceedings of the Eighth Conference on Computational Natural Language ...
- The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages**
J Hajic, M Ciaramita, R Johansson, D Kawahara, MA Martí, L Màrquez, ...
Proceedings of the Thirteenth Conference on Computational Natural Language ...
- The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies**
M Surdeanu, R Johansson, A Meyers, L Màrquez, J Nivre
CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural ...
- Boosting trees for anti-spam email filtering**
X Carreras, L Marquez
arXiv preprint cs/0109015
- SVMTool: A general POS tagger generator based on Support Vector Machines**
J Giménez, L Marquez
In Proceedings of the 4th International Conference on Language Resources and ...

TITLE	CITED BY	YEAR
Introduction to the CoNLL-2005 shared task: Semantic role labeling	1001	2005
Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling	1004	2004
The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages	629	2009
The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies	590	2008
Boosting trees for anti-spam email filtering	580	2001
SVMTool: A general POS tagger generator based on Support Vector Machines	516	2004



Thanks! / Take-home messages

- Achieve **community-wide advancements** through the establishments of **openly shared experimental settings** (data, metrics) through *competition*
- “Cement” the results into **openly available publications and code**
- Make it appealing through **visibility rewards** (most prominently, citations)
- Need to be aware of **limitations!**
Example: “*rat-race chasing numbers*”

Simone Ponzetto / Data Science in Action

18.11.2022

