How to win a shared task

Rob van der Goot

Who am I?

- Bachelor, master, PhD from University of Groningen
- Postdoc and assistant professor at the IT University of Copenhagen
- Research on:
  - Parsing and normalization for social media data
  - Multi-task learning
  - Data-scarce setups (low-resource datasets)
  - Question assumptions

Warning

- Subjectivity
- Biased view
- Bragging
- Criticizing others
- Advertising my own tool

Why should you trust me on this topic?

- Third place Dutch championships volleyball 2008
- 2nd best Jumbo of NL (2011)
- Participant Guinnes record of most people (361) walking 1,000 meter on ice barefoot (2013)
- 3th place SemEval 2014 task1
- 6th place SemEval 2015 task2
- 1st place CLIN26 2015
- 1st/2nd place WNUT 2020 shared task
- Outstanding paper award EACL 2021
- Outstanding reviewer ACL 2021
- 3th and last place WNUT 2021 shared task
- Best paper award WNUT 2022
- 3th-56th place SemEval 2022

Shared tasks

- Who is the winner?
- How to rank high?
- Explainable Detection of Online Sexism
- How (not) to design a poster

Shared tasks

Who is a winner?

MultiLexNorm: A Shared Task on Multilingual Lexical Normalization

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli and Wladimir Sidorenko





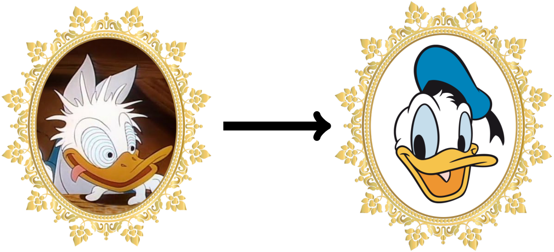







Lexical Normalization

Lexical normalization is the task of transforming an utterance into its standard form, word by word, including both one-to-many (1-n) and many-to-one (n-1) replacements.



Lexical Normalization

State before shared task:

- Most work on English
- Also work on single other languages
- Varieties in task definitions, guidelines and metrics
- No common evaluation benchmark

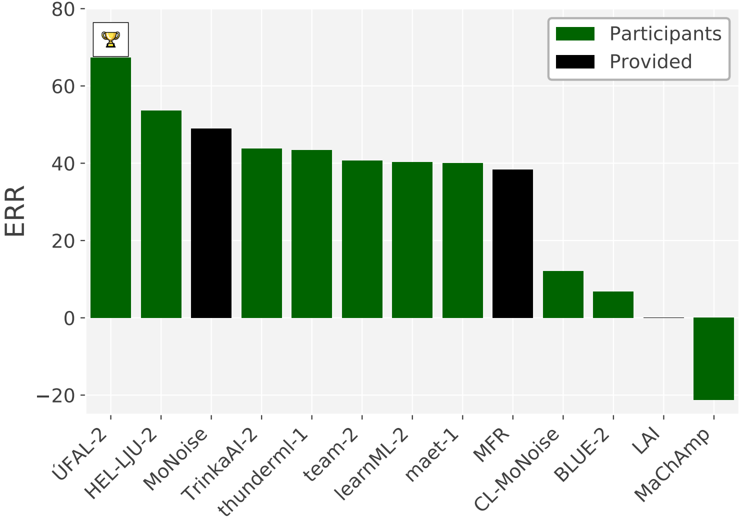
MultiLexNorm

- Combination of existing datasets
- Annotation style and file format converged
- “new” evaluation metric
- External evaluation (UD)

MultiLexNorm

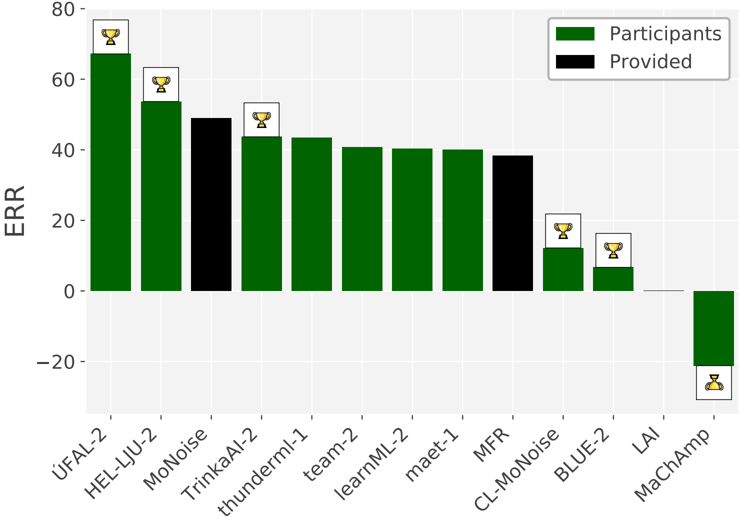
Lang.	Language name	Normalization example
DA	Danish	De skarpe lamper gjorde destromindre ek bedre .
		De skarpe lamper gjorde destro mindre ikke bedre .
DE	German	ogäj isch hätts auch dwiddern könn
		Okay ich hätte es auch twittern können
EN	English	u hve to let ppl decide what dey want to do
		you have to let people decide what they want to do
ES	Spanish	@username cuuxamee sii peroo veen yaa eem
		@username escúchame sí pero ven ya eh
HR	Croatian	svi frendovi mi nešto rade , veceras san osta sam .
		svi frendovi mi nešto rade , večeras sam ostao sam .
ID-EN	Indonesian-English	pdhal not fully bcs those ppl jg sih .
		padahal not fully because those people juga sih .
IT	Italian	a Roma è cosí primavera che sembra gia giov
		a Roma è così primavera che sembra già giovedì
NL	Dutch	Kga me wss trg rolle vant lachn
		Ik ga me waarschijnlijk terug rollen van het lachen
SL	Slovenian	jst bi tud najdu kovanec vreden veliko denarja .
		jaz bi tudi našel kovanec vreden veliko denarja .
SR	Serbian	komunalci kace pocne kaznjavanje ?
		komunalci kad počne kažnjavanje ?
TR	Turkish	He o dediyin suala cvb verdim
		He o dediğin suale cevap verdim
TR-DE	Turkish-German	@username Yerimm senii , damkee schatzymm :-*
		@username Yerim seni , danke Schatzym :-*

Results



80
60
40
20
0
−20
ERR
Participants
Provided
ÚFAL-2
HEL-LJU-2
MoNoise
TrinkaAI-2
thunderml-1
team-2
learnML-2
maet-1
MFR
CL-MoNoise
BLUE-2
LAI
MaChAmp

Results



80
60
40
20
0
−20
ERR
Participants
Provided
ÚFAL-2
HEL-LJU-2
MoNoise
TrinkaAI-2
thunderml-1
team-2
learnML-2
maet-1
MFR
CL-MoNoise
BLUE-2
LAI
MaChAmp

MultiLexNorm

- First to use transformers seq2seq (char level)
- First to do external evaluation on sentiment analysis, hatespeech
- First multi-lingual models
- First cross-lingual models
- First to model the task as sequence labeling
- Many new methods for generating training data

My most important advice; make sure you have an interesting research question!

My most important advice; make sure you have an interesting research question!

- Note that you are also dependent on data and organization for the ranking!
  - Data quality can be questionable
  - Metrics can be wrongly implemented/chosen

My most important advice; make sure you have an interesting research question!

- Note that you are also dependent on data and organization for the ranking!
  - Data quality can be questionable
  - Metrics can be wrongly implemented/chosen
- Convinced yet?

How to rank first though?

SemEval 2014

SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment

SemEval 2014

SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment

- Johan Bos focused on the RTE part
- I focused on STS

SemEval 2014

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

Table 1: Examples of sentence pairs with their gold relatedness scores (on a 5-point rating scale).

SemEval 2014

It was 2014:

- Word embeddings
- Feature-based systems

SemEval 2014

Table 2: Pearson correlation and MSE obtained on the test set for each feature group in isolation.

Feature group	**p [−PPDB]**	**p [+PPDB]**	**MSE [−PPDB]**	**MSE [+PPDB]**
Logical model	0.649	0.737	0.590	0.476
Noun/verb overlap	0.647	0.676	0.592	0.553
DRS	0.634	0.667	0.610	0.569
Wordnet novelty	0.652	0.651	0.590	0.591
RTE	0.621	0.620	0.626	0.627
CDSM	0.608	0.609	0.681	0.679
IDs	0.493	0.493	0.807	0.807
Synset	0.414	0.417	0.891	0.889
Word overlap	0.271	0.340	0.944	0.902
Sentence length	0.227	0.228	0.971	0.971
All with IDs	0.836	0.842	0.308	0.297
All without IDs	0.819	**0.827**	0.336	**0.322**

SemEval 2014

Table 2: Pearson correlation and MSE obtained on the test set for each feature group in isolation.

Feature group	p [−PPDB]	p [+PPDB]	MSE [−PPDB]	MSE [+PPDB]
Logical model	0.649	0.737	0.590	0.476
Noun/verb overlap	0.647	0.676	0.592	0.553
DRS	0.634	0.667	0.610	0.569
Wordnet novelty	0.652	0.651	0.590	0.591
RTE	0.621	0.620	0.626	0.627
CDSM	0.608	0.609	0.681	0.679
IDs	0.493	0.493	0.807	0.807
Synset	0.414	0.417	0.891	0.889
Word overlap	0.271	0.340	0.944	0.902
Sentence length	0.227	0.228	0.971	0.971
All with IDs	0.836	0.842	0.308	0.297
All without IDs	0.819	**0.827**	0.336	**0.322**

SemEval 2014

ID	Compose	r	ρ	MSE
ECNU_run1	S	0.828	0.769	0.325
StanfordNLP_run5	S	0.827	0.756	0.323
The_Meaning_Factory_run1	S	0.827	0.772	0.322
UNAL-NLP_run1		0.804	0.746	0.359
Illinois-LH_run1	P/S	0.799	0.754	0.369
CECL_ALL_run1		0.780	0.732	0.398
SemantiKLUE_run1		0.780	0.736	0.403
RTM-DCU_run1		0.764	0.688	0.429
UTexas_run1	P/S	0.714	0.674	0.499
UoW_run1		0.711	0.679	0.511
FBK-TR_run3	P	0.709	0.644	0.591
BUAP_run1	P	0.697	0.645	0.528
UANLPCourse_run2	S	0.693	0.603	0.542
UQeResearch_run1		0.642	0.626	0.822
ASAP_run1	P	0.628	0.597	0.662
Yamraj_run1		0.535	0.536	2.665
asjai_run5	S	0.479	0.461	1.104

Table 7: Primary run results for the relatedness subtask (r for Pearson and ρ for Spearman correlation). The table also shows whether a system exploits composition information at either the phrase (P) or sentence (S) level.

SemEval 2022

- The first one to participate in > 2 SemEval tasks (more details on Friday!)
- Didn't rank first on any of them
- Why?

Limited time, so I:

- used mBERT and RemBERT
- didn't find additional data
- didn't tune hyperparameters
- didn't do synthetic data generation
- didn't tune pre-/post- processing
- didn't use sentence ID's as feature

Which LM to use?

`https://www.youtube.com/watch?v=MXGCNZvqS1o`

Which LM to use?

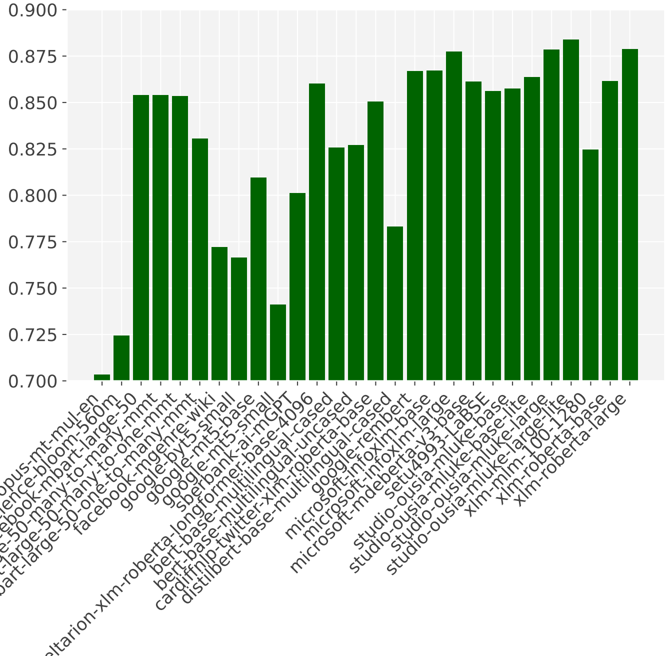
- Paper: Evidence > Intuition: Transferability Estimation for Encoder Selection. Elisa Bassignana, Max Müller-Eberstein, Mike Zhang, Barbara Plank. 2022 EMNLP
- I bruteforced it for 2 common benchmarks: `https://robvanderg.github.io/blog/tune_lms.htm`

Which LM to use?

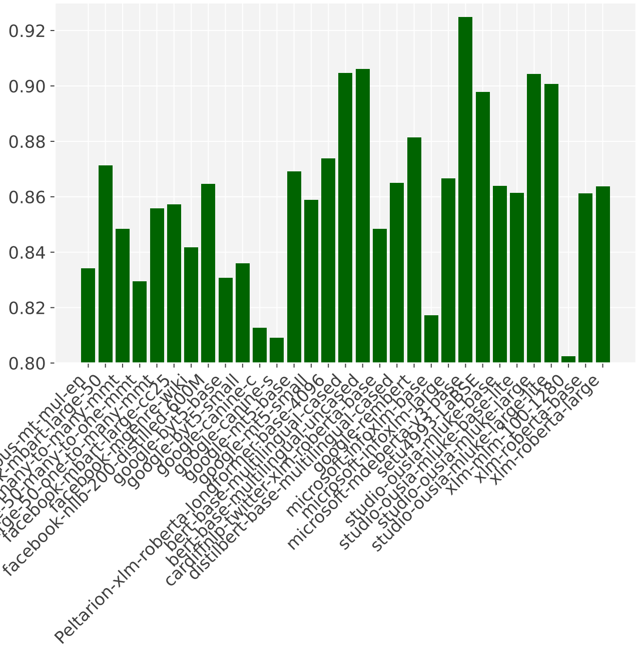
Setup:

- Benchmarks: UD and Glue (subsets)
- Model: MaChAmp
- MLMs: all multilingual (>10 languages) language models

Which LM to use? (UD)



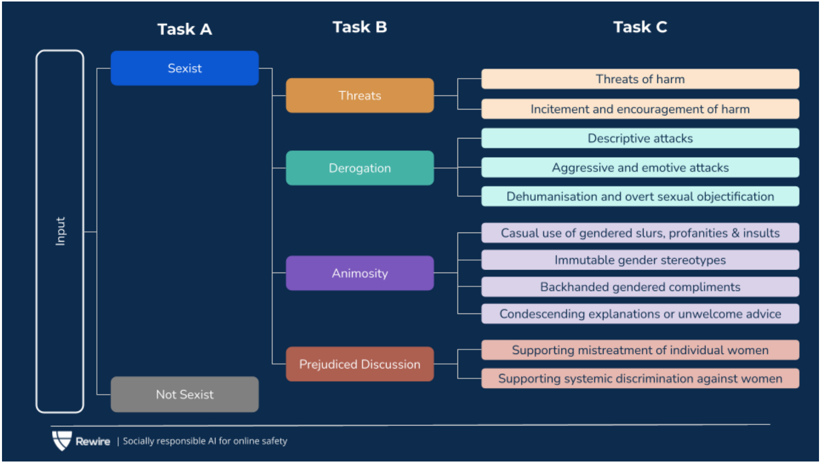
Which LM to use? (GLUE)



Which LM to use?

- Was this interesting?
- Was it costly?
- Can it help you win?

SemEval 2023 Task10: Explainable Detection of Online Sexism



- Macro-F1 for all subtasks?
- Assume gold binary detection for next steps?

SemEval 2023 Task10

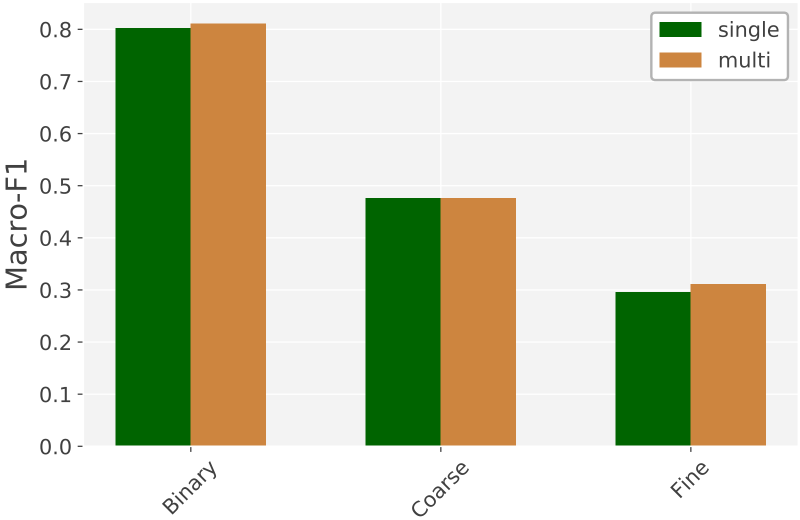
Framework: MaChAmp

Massive Choice, Ample Tasks (MACHAMP):
A Toolkit for Multi-task Learning in NLP

Rob van der Goot Ahmet Üstün Alan Ramponi Ibrahim Sharaf
Barbara Plank
IT University of Copenhagen University of Groningen University of Trento
Fondazione the Microsoft Research - University of Trento COSBI Factmata
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it
ibrahim.sharaf@factmata.com, bapl@itu.dk

- Default parameters
- Each task separate or one joint model?
- Which LM?
- Language modeling
- Hierarchical separate models

SemEval 2023 Task10



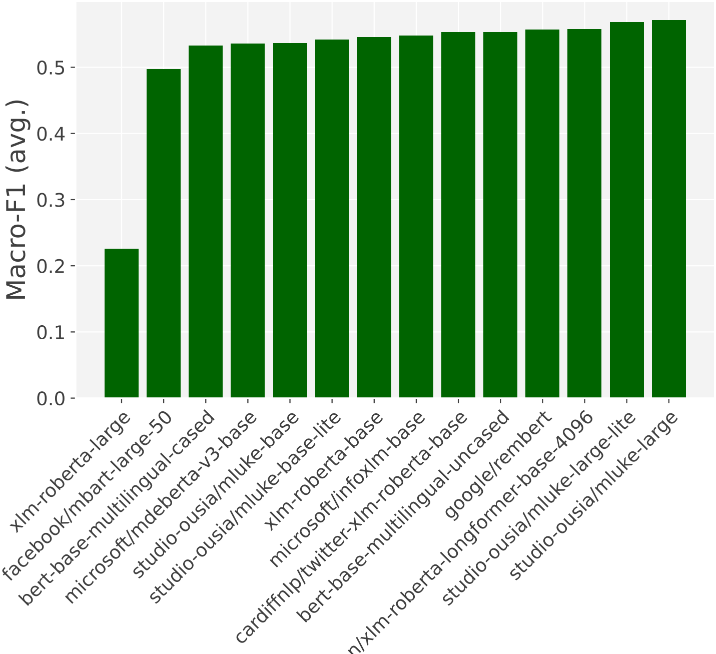
Multi-task learning in MaChAmp

```
{
    "task10_1": {
        "train_data_path": "data/task10/train.conll",
        "dev_data_path": "data/task10/dev.conll",
        "sent_idxs": [
            0
        ],
        "tasks": {
            "sexism": {
                "task_type": "classification",
                "column_idx": 1,
                "metric": "f1_macro"
            }
        }
}
```

Multi-task learning in MaChAmp

```
python3 train.py --dataset_configs configs/task10_1.json
python3 train.py --dataset_configs configs/task10_2.json
python3 train.py --dataset_configs configs/task10_3.json
python3 train.py --dataset_configs configs/task10_1.json\
 configs/task10_2.json configs/task10_3.json
```

SemEval 2023 Task10



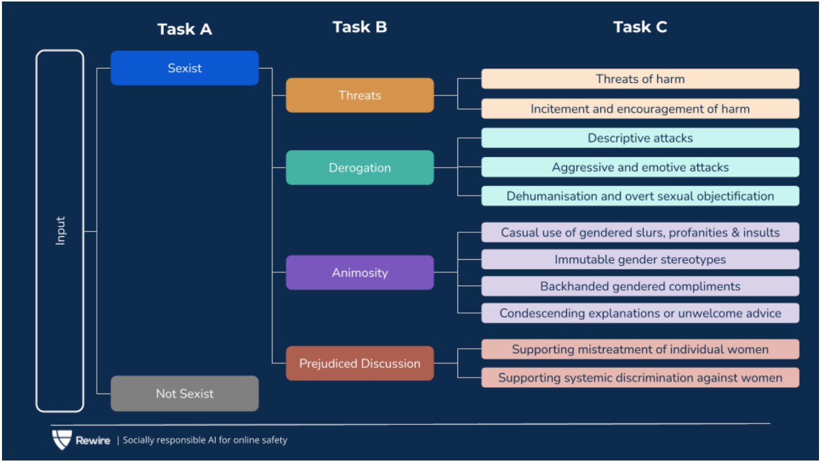
SemEval 2023 Task10

Subtask	Multi
Binary	82.68
Coarse	55.91
Fine	33.11

Table: Macro-f1 scores of mLUKE-large

SemEval 2023 Task10

A classifier per decision:



SemEval 2023 Task10

Subtask	Multi	Hierarchical
Binary	82.68	82.68
Coarse	55.91	53.39
Fine	33.11	31.30

Table: Macro-f1 scores of mLUKE-large

SemEval 2023 Task10

Improvements

- Allow more instances to go through (confidence?)
- Tune each model

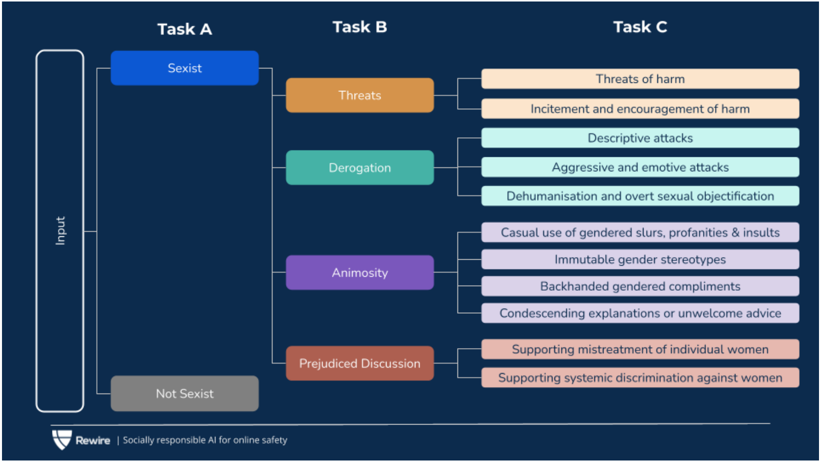
SemEval 2023 Task10

Improvements

- Allow more instances to go through (confidence?)
- Tune each model
- Do it the other way around: Fine-grained informs the others

SemEval 2023 Task10

Can we do it the other way around?



SemEval 2023 Task10

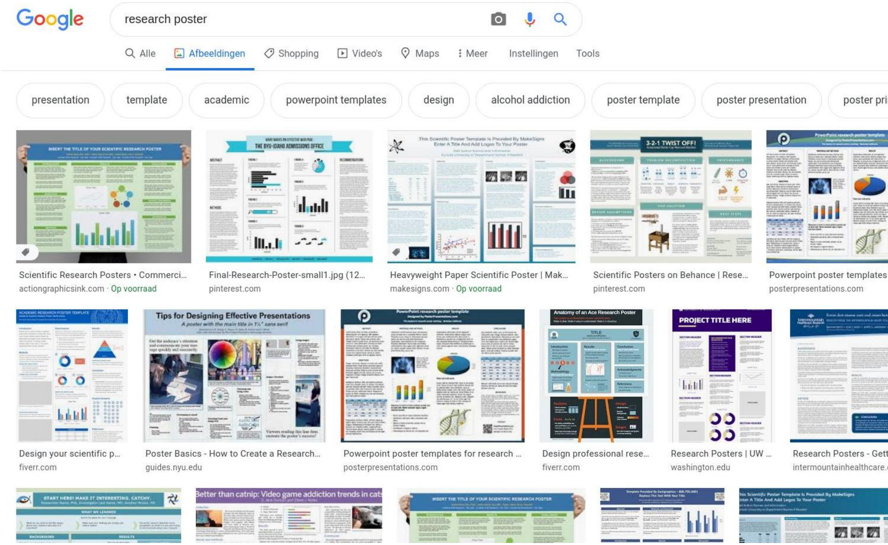
Model	Score
Binary	82.68
Coarse	83.43
Fine	83.48

Table: Macro-f1 scores of subtask predictions on binary task

SemEval 2023 Task10

Good luck! may we all win

How to design a poster



How to design a poster

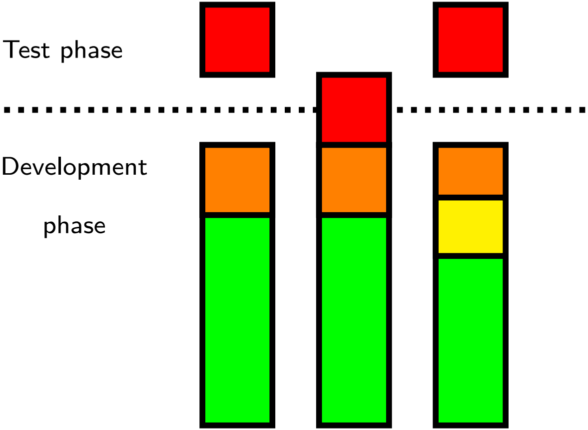
- You already wrote a paper, don't have to make another one
- You spend a lot of time on the project, take a couple of hours to present it well (its the fun part!)
- People will see dozens or even hundreds of presentations
- Prioritize visual clues and simple take away messages!

How to design a poster

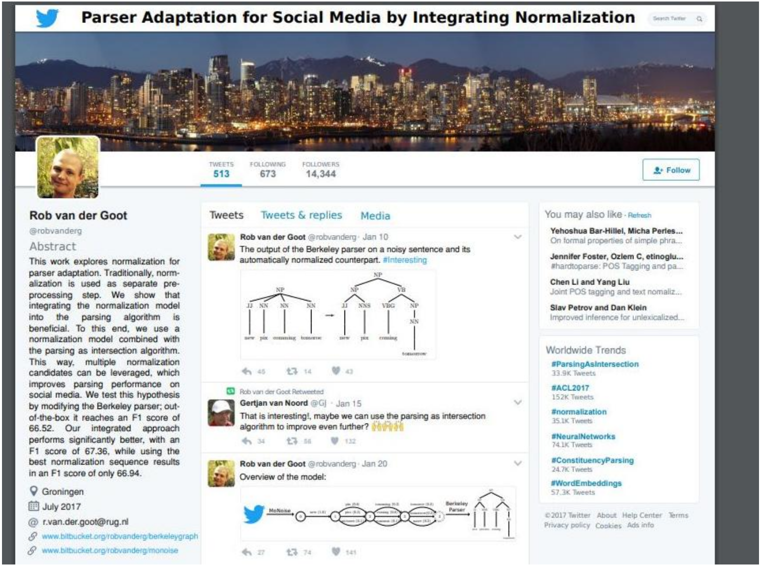
The following examples are completely cherry picked from my own collection

We Need to Talk About train-dev-test Splits

Rob van der Goot



- Pick a design or visualization that people will remember


Parser Adaptation for Social Media by Integrating Normalization
TWEETS 513
FOLLOWING 673
FOLLOWERS 14,344
Follow
Rob van der Goot
@robvanderg
Abstract
This work explores normalization for parser adaptation. Traditionally, normalization is used as separate pre-processing step. We show that integrating the normalization model into the parsing algorithm is beneficial. To this end, we use a normalization model combined with the parsing as intersection algorithm. This way, multiple normalization candidates can be leveraged, which improves parsing performance on social media. We test this hypothesis by modifying the Berkeley parser; out-of-the-box it reaches an F1 score of 66.52. Our integrated approach performs significantly better, with an F1 score of 67.36, while using the best normalization sequence results in an F1 score of only 66.94.
Groningen
July 2017
r.van.der.goot@rug.nl
www.bitbucket.org/robvanderg/berkeleygraph
www.bitbucket.org/robvanderg/monoise
Tweets
Tweets & replies
Media
Rob van der Goot @robvanderg · Jan 10
The output of the Berkeley parser on a noisy sentence and its automatically normalized counterpart. #Interesting
NP
JJ NN NN NN
new pix comming tomoroe
NP
NP VB
JJ NNS VBG NP
NN
new pix coming tomorrow
45 14 43
Rob van der Goot Retweeted
Gertjan van Noord @GJ · Jan 15
That is interesting!, maybe we can use the parsing as intersection algorithm to improve even further?
34 56 132
Rob van der Goot @robvanderg · Jan 20
Overview of the model:
MoNoise
Berkeley Parser
27 74 141
You may also like · Refresh
Yehoshua Bar-Hillel, Micha Perles...
On formal properties of simple phra...
Jennifer Foster, Ozlem C, etinoglu...
#hardtoparse: POS Tagging and pa...
Chen Li and Yang Liu
Joint POS tagging and text nomaliz...
Slav Petrov and Dan Klein
Improved inference for unlexicalized...
Worldwide Trends
#ParsingAsIntersection
33.9K Tweets
#ACL2017
152K Tweets
#normalization
35.1K Tweets
#NeuralNetworks
74.1K Tweets
#ConstituencyParsing
24.7K Tweets
#WordEmbeddings
57.3K Tweets
©2017 Twitter About Help Center Terms
Privacy policy Cookies Ads info

MaChAmp #068

Attacks:

seq	classification
seq-bio	dependency
seq-multi	mlm
string2string	seq2seq

MaChAmp is a multi-task NLP toolkit, it can seemingly effortlessly handle multiple NLP tasks simultaneously. It has functionality for joint training, continuous training, dataset smoothing, loss weights and dataset embeddings.

More information on: machamp-nlp.github.io

Notes

How to use MaChAmp.

This is what the dataset configuration file looks like:

```
{"UD": {,
   "train_data_path": "data/ewt.train",
   "validation_data_path": "data/ewt.dev",
   "word_idx": 1,
   "tasks": {
     "lemma": {
       "task_type": "string2string",
       "column_idx": 2,
     }
   }
 }
```

Then I can train with the following command:

```
python3 train.py --dataset_config ewt.json
```

And predict with:

```
python3 predict.py logs/ewt/model.tar.gz
   data/ewt.dev preds/ewt.dev.out
```

How to design a poster

- Pick a design that people will remember
  - Not always easy: think of an interesting, nice, or funny example in your data

Bleaching Text: Abstract Features for Cross-lingual Gender Prediction

Rob van der Goot[♡] **Nikola Ljubešić**[♠] **Ian Matroos**[♡] **Malvina Nissim**[♡] **Barbara Plank**[♡♣]
[♡] Center for Language and Cognition, University of Groningen, The Netherlands
[♠] Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
[♣] IT University of Copenhagen, Copenhagen, Denmark
{r.van.der.goot,i.matroos,m.nissim}@rug.nl, nljubesi@gmail.com, bplank@itu.dk


THICK BLEACH

Lexical Normalization for Code-switched Data and its Effect on POS Tagging

IT UNIVERSITY OF CPH

Rob van der Goot and Özlem Çetinoğlu

benimde saprachdiplom vardi ama yinede gittim kursa

????

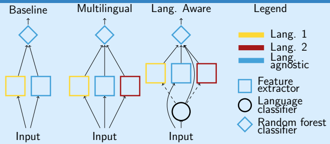
Benim de Sprachdiplom vardı ama yine de gittim kursa

Ahh!

Data

Raw:	@Erkan1903 nerdee 3 **semester**dayim dha.					
Tok+Anon:	@username	nerdee	3	**semester**dayim	dha	.
Norm	@username	Nerde	3.	**Semester**dayım	daha	.
	OTHER	TR	OTHER	MIXED	TR	OTHER
Seg+CS:	@username Nerde 3. **Semester**§da -yım daha .					

Models



Contributions

- We introduce a publicly available dataset for Tr-De with normalization, language ID and POS layers
- Publicly available normalization models for multiple languages without language-specific heuristics
- Reach new SOTA for normalization on code-switched data
- Show that normalization is beneficial for POS tagging

Code: `https://bitbucket.org/robvanderg/csmonoise`
Data: `https://github.com/ozlemcek/TrDeNormData`

Results

Model	Normalization Id-En	Normalization Tr-De	POS Tr-De
LAI	74.03	67.02	60.77
Monolingual (Id/De)	*94.62	76.33	*63.47
Multilingual	94.27	*78.28	*64.06
Language-aware	94.32	77.83	*63.92
Gold	*100.00	*100.00	*67.75

Lexical Normalization for Code-switched Data and its Effect on POS Tagging

IT UNIVERSITY OF CPH

Rob van der Goot and Özlem Çetinoğlu

benimde saprachdiplom vardi ama yinede gittim kursa

????

Benim de Sprachdiplom vardı ama yine de gittim kursa

Ahh! 👍

Contributions

Data

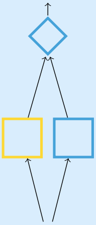
Raw:

Tok+Anon:

Norm

Seg+CS:

Models



R

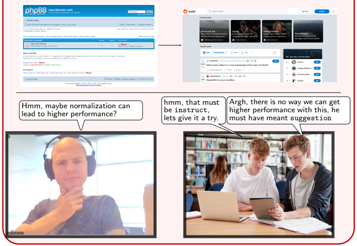
Increasing Robustness for Cross-domain Dialogue Act Classification on Social Media Data

Marcus Vielsted, Nikolaj Wallenius, and Rob van der Goot

Task

Label	Example
propositionalQuestion	“r u serious?”
setQuestion	“what list should i put him in?”
choiceQuestion	“shaken or stirred?”
inform	“i wonna chat”
elaborate	“and dr phil said so.”
continuer	“I know, but it threw me”
agreement	“i agree”
disagreement	“no, I didnt even look.”
correction	“i meant to write the word may.”
greeting	”hey ladies”
goodbye	“see u all laters”
positiveExpression	“yay!”
negativeExpression	“ewwwwww lol”
offer	“il get you a cheap flight to hell:)?”
suggestion	“We should have a club”
instruct	“shut the fuck up.”
acceptAction	“yeah i should toss it”
declineAction	“i don’t wanna”
misc	:tongue:

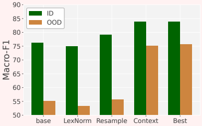
Problem



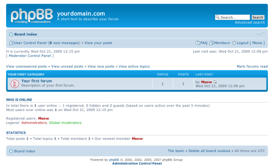
Results

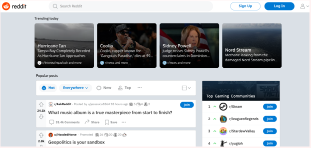
Split	ID	OOD
train	4,800	—
dev	600	853
test	600	852

Sizes of the annotated parts of NPSChat (ID) and Reddit (OOD)



Problem





Hmm, maybe normalization can lead to higher performance?

hmm, that must be `instruct`, lets give it a try.

Argh, there is no way we can get higher performance with this, he must have meant `suggestion`

Take-away Messages

- **Have an interesting research question!**
- Have fun trying motivated approaches and arbitrary changes to improve performance
- Your poster is not a paper