
ROB AND THE CHALLENGES OF
ROBUSTNESS IN NLP

Today

Robustness through

- Lexical Normalization
- Multi-task learning

Lexical Normalization

u	hve	to	let	ppl	decide	what	dey	want	to	do
you	have	to	let	people	decide	what	they	want	to	do

Lexical Normalization

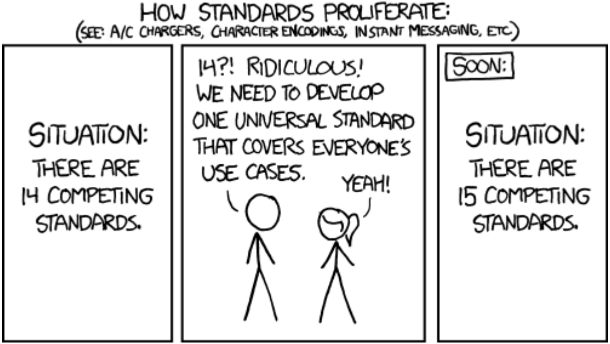
Situation in 2015:

- some benchmarks for English: main one LexNorm
- Some people working on their own languages
- Differences in models, task definitions and metrics

Lexical Normalization

Situation in 2019:

- First model that works for multiple languages (7): MoNoise
- SOTA on all evaluated languages
- Proposed a new metric: ERR



Lexical Normalization

Situation in 2019:

- First model that works for multiple languages (7): MoNoise
- SOTA on all evaluated languages
- Proposed a new metric

Corpus	Lang	ERR	Precision	Recall	Prev. SOTA	Metric	Prev.	MoNoise
GhentNorm	NL	44.62	89.19	50.77	Schulz et al. (2016)	WER	3.2	1.36[5]
TweetNorm	ES	38.73	94.37	41.19	Porta and Sancho (2013)	OOV-Precision	63.4	70.40
LexNorm1.2	EN	59.21	80.87	77.56	Li and Liu (2015)	OOV Accuracy	87.58	87.63
LexNorm2015	EN	77.09	95.49	80.91	Jin (2015)	F1	84.21	86.58
IWT	TR	28.94	96.24	30.12	Eryiğit et al. (2017)	OOV Accuracy	67.37	48.99
Janes-Norm	SL	31.67	85.19	0.3833	Ljubešic et al. (2016) L1	CER	0.38	0.53
Janes-Norm	SL	63.90	95.66	0.6694	Ljubešic et al. (2016) L3	CER	1.58	2.24
ReLDI-hr	HR	51.65	95.66	0.541				
ReLDI-sr	SR	64.61	94.70	68.43				

Lexical Normalization

Situation in 2021:

- Nothing changed

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












MultiLexNorm

Introduced in a shared task (WNUT):

- 12 languages
- annotation style and file format converged
- ERR is main metric
- Downstream evaluation on dependency parsing on 7 treebanks

MultiLexNorm

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.

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 :-*

Metric

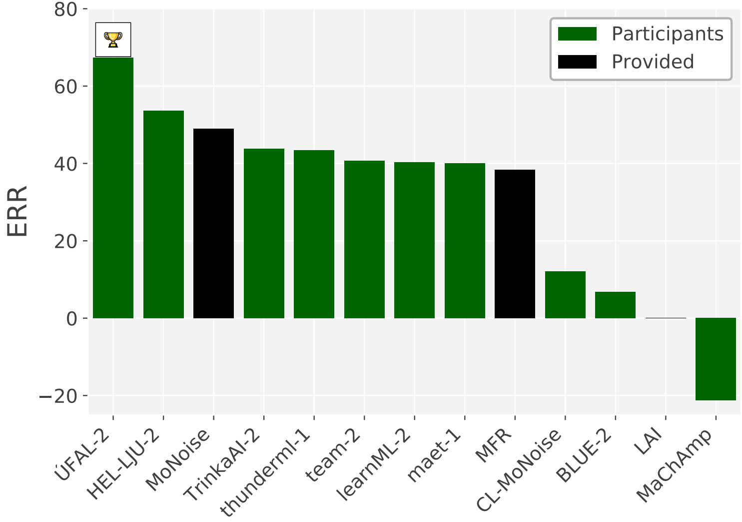
- Previously: accuracy, accuracy over OOV words, F1 score, BLEU, word error rate, character error rate, etc.
- Now: accuracy normalized for amount of words to be normalized. Error Reduction Rate:

$$ERR = \frac{\%\text{accuracy} - \%\text{words\_not\_normed}}{100 - \%\text{words\_not\_normed}}$$

MutliLexNorm

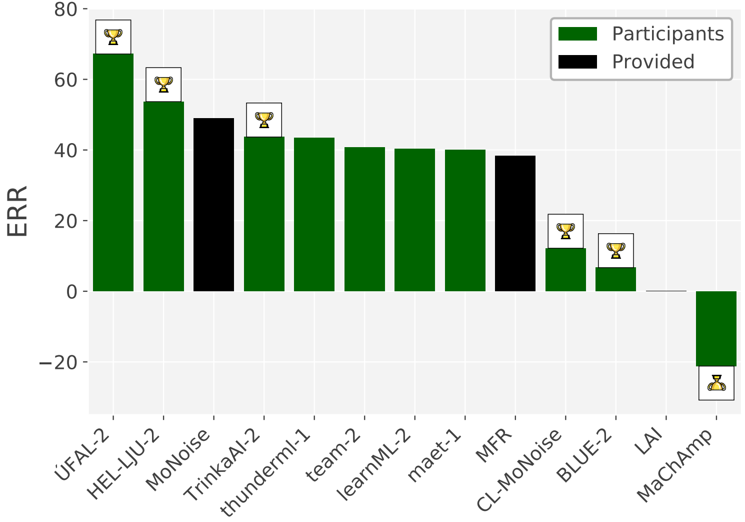
- ÚFAL: ByT5 for every word; synthetic data
- HEL-LJU: Pre-classify type of normalization (BERT) $\mapsto$ Char-SMT
- MoNoise: Feature-based, generate candidates and rank
- BLUE: NMT MBart-50
- CL-MoNise: Cross-lingual
- MaChAmp: normalization as sequence labeling

Results



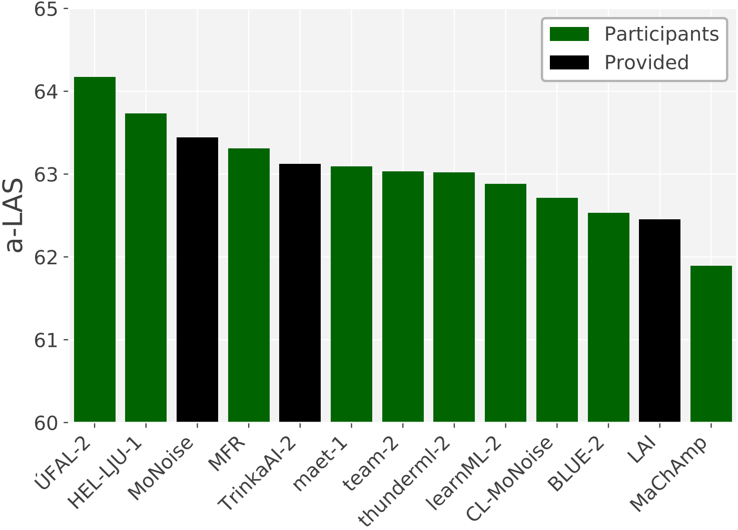
80
60
40
20
0
−20
ERR
Participants
Provided
ÚFAL-2
HEL-LJU-2
MoNoise
TrinkaAI-2
thunderml-1
team-2
learnML-2
maet-1
MFR
CL-MoNoise
BLUE-2
LAI
MaChAmp

Results

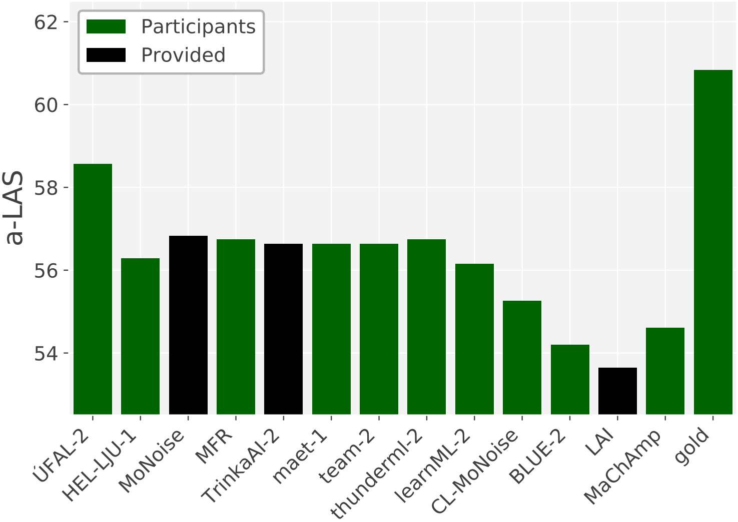


ERR
80
60
40
20
0
−20
Participants
Provided
ÚFAL-2
HEL-LJU-2
MoNoise
TrinkaAI-2
thunderml-1
team-2
learnML-2
maet-1
MFR
CL-MoNoise
BLUE-2
LAI
MaChAmp

Extrinsic Evaluation (avg.)



Extrinsic Evaluation (EN-MoNoise)



Findings

- Include detection in task
- Multi-lingual benchmark
- Wide variety of models
- Near-human performance

Open problems

- Cross-lingual/multi-lingual normalization
- Tokenization
- Limited downstream gains; lexical level might not be enough
- Bias in languages
- Bias in data source

Multi-task learning

- xSID: auxiliary tasks
- MaChAmp at SemEval 2022 and 2023: Intermediate training

Framework

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

xSID: Cross-lingual Slot and Intent Detection

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanovìc, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi and Barbara Plank









Slot and Intent Detection

I'd like to see the showtimes for Silly Movie 2.0 at the movie house
Intent: SearchScreeningEvent

xSID

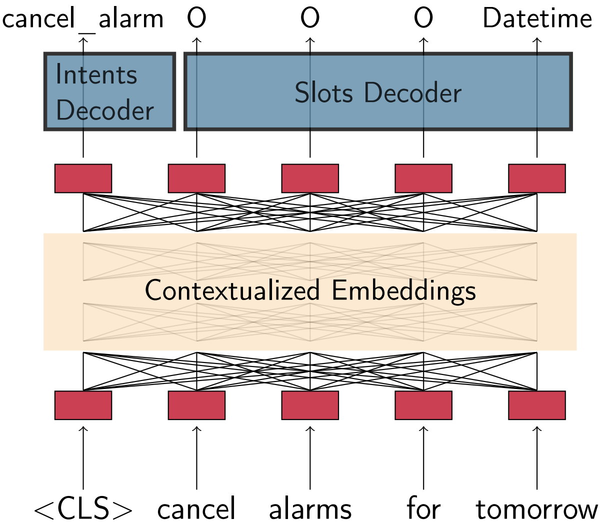
ar	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	映画館のSilly Movie 2.0の上映時間を見せて。
kk	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын көргім келеді
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Silly Movie 2.0'ın sinema salonundaki seanslarını görmek istiyorum
zh	我想看Silly Movie 2.0 在影院 的放映

Experiments

Baselines

- Baseline: contextualized embeddings with joint intent+slots
- Stronger baseline: translate training data to target language and map slot labels with attention (NMT-TRANSFER)

Experiments



Experiments

New models:

- Train on auxiliary task in target language:
  - Masked language modeling (AUX-MLM)
  - Neural machine translation (AUX-NMT)
  - UD-parsing (AUX-UD)

Experiments

Evaluate 2 embeddings

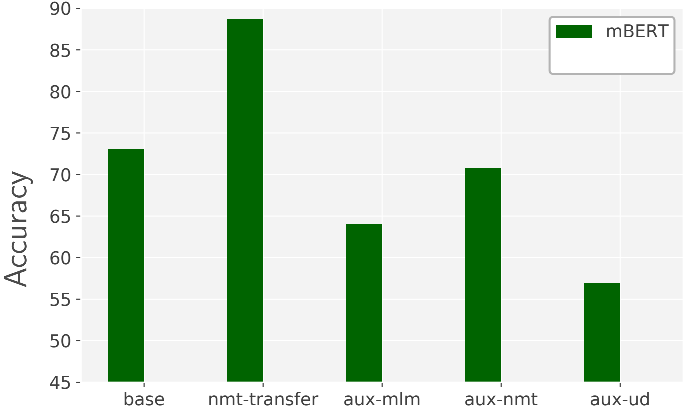
- mBERT: trained on 104 languages (12/13)
- XLM15: trained on 15 languages (5/13)

Results

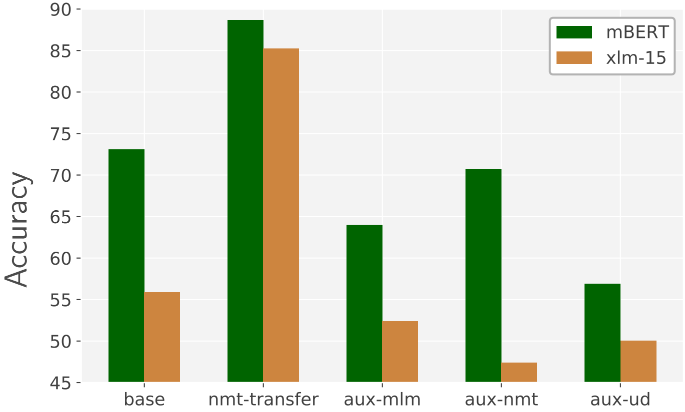
model	Time (minutes)
base	46
nmt-transfer	5,213
aux-mlm	193
aux-nmt	373
aux-ud	79

Table: Average minutes to train a model, averaged over all languages and both embeddings. For nmt-transfer we include the training of the NMT model.

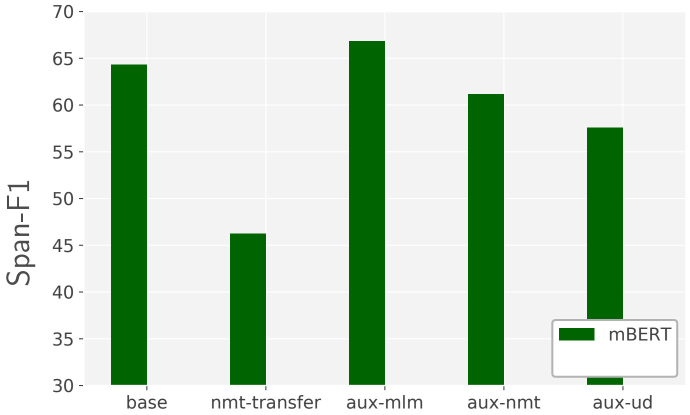
Results (intents)



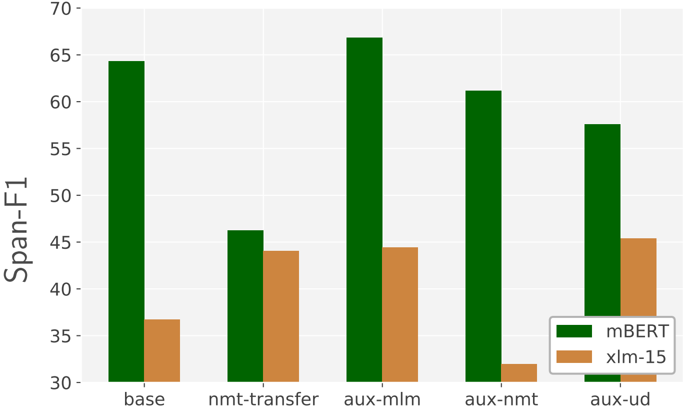
Results (intents)



Results (slots)



Results (slots)



Resolved mysteries

Sentence level:

- NMT-transfer is hard to outperform, but costly
- Even baseline hard to beat

Span level:

- NMT-transfer performs bad (due to alignment)
- In-LM languages: only MLM helps
- Out-LM languages: More explicit tasks (UD) are faster and lead to better performance

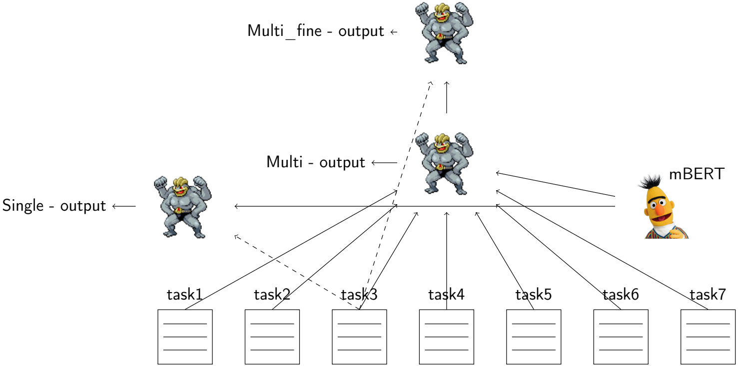
Unresolved mysteries

- Can NMT be used as auxiliary task?
- Are there better sentence level auxiliary tasks?
- Can NMT-transfer be improved with better word alignment?
- NMT and MLM hyperparameters
- Modeling jointly versus sequentially

Extensions

- SID4LR
    - Neapolitan
    - Swiss German
- More coming!

A newer multi-task setup: Intermediate task finetuning



Other names:

- Task Adaptive PreTraining (TAPT)
- Pre-finetune
- Multi-task finetuning
- Multi-task prompted training
- Supplementary training on intermediate labeled data tasks (STILT)
- Intermediate task finetuning
- Intermediate task training
- Intertraining
- ...

Intermediate task finetuning

- STILT
- T0
- Ext5
- MUPPET
- In-BoXBART
- Sem-mmmBERT
- ...

Intermediate task finetuning

- STILT
- T0
- Ext5
- MUPPET
- In-BoXBART
- Sem-mmmBERT
- ...

MaChAmp at SemEval-2022 Tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task Multi-lingual Learning for a Pre-selected Set of Semantic Datasets

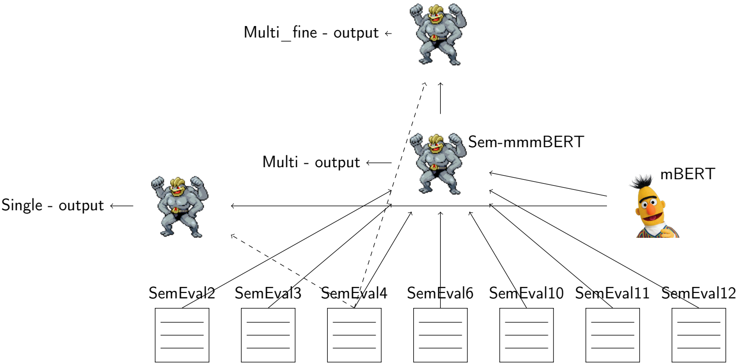
Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Intermediate task finetuning

Research questions:

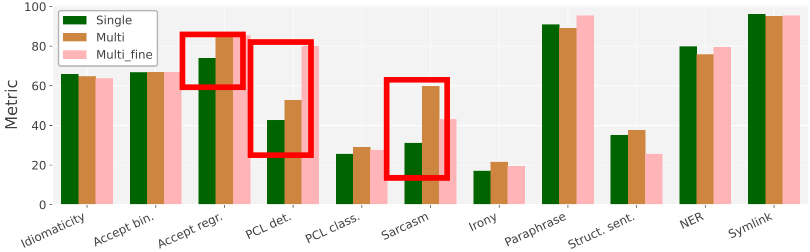
- Can we use this approach in an autoencoder language model?
- Is intermediate task finetuning also beneficial for a somewhat arbitrary set of semantic tasks?

Intermediate task finetuning



SemEval Task	Included sub-tasks	Languages
2: Multilingual Idiomaticity Detection	Idiomaticity detection (1-shot)	EN, PT, GL
3: PreTENS	1: Binary acceptability	EN, IT, FR
	2: Regression acceptability	EN, IT, FR
4: Patronizing and Condescending Language Detection	1: Binary PCL detection	EN
	2: Multi-label PCL classification	EN
6: iSarcasmEval	1: Sarcasm detection	EN, AR
	2: Irony-labeling	EN
	3: Paraphrase sarcasm detection	EN, AR
10: Structured Sentiment Analysis	Expressions, entities and relations	CA, EN, ES, EU, NO
11: MultiCoNER - Multilingual Complex Named Entity Recognition	Named Entity Recognition	BN, DE, EN, ES, FA, HI, KO, MI, NL, RU, TR, ZH
12: Symlink	Entities and relations	EN

Intermediate task finetuning



Resolved mysteries

- Medium performance baseline tgt task
- Largest gains for some sub-tasks (task-relatedness)
- Language

Unresolved mysteries

- How can we do better?
  - Use other LM's
  - Finetune hyperparameters
  - Add/select pre-training tasks
- Can we predict which tasks to select?
  - Hard without many overlapping datasets (task/language dimension)
  - Too many combinations possible

- Many new bencharks; (almost) all publicly available
- MaChAmp; easy SOTA for many NLP tasks
- New focus: simple tasks in challenging setups